DETECCION DE CÁNCER DE MAMA MEDIANTE MACHINE LEARNING Y CNN

BREAST CANCER DETECTION USING MACHINE LEARNING AND CNN

Alumno: Felipe Sebastián Galarza
Email: 9001649@alumnos.ufv.es
Grado en Business Analytics
Curso académico 2024-2025
Tutora: María Jesús Gómez Fernández
Facultad: Derecho, Empresa y Gobierno

Resumen

El cáncer de mama es una de las enfermedades con mayor repercusión en la población femenina a nivel mundial y para disminuir su riesgo es fundamental detectarlo de forma temprana. Para ello este trabajo tiene el objetivo de crear un sistema de apoyo al diagnóstico a partir características numéricas e imágenes ultrasonidos de los tumores. Se busca clasificarlos en benignos o malignos mediante el uso de modelos de Machine Learning, para las características numéricas, y Redes Neuronales Convolucionales (CNN), para las imágenes ultrasonidos. En el primer caso se comparan, a través de distintas métricas de evaluación, los modelos Regresión Logística, Random Forest y Gradient Boosting. Entre ellos el que mejor rendimiento presenta, teniendo en cuenta la cantidad limitada de datos que se tiene, es la Regresión Logística porque es el que presenta la mayor precisión y generalización en datos nuevos. Para el segundo caso, debido a la falta de capacidad computacional, se utiliza el modelo preentrenado DenseNet121 mediante la técnica de Partial Fine Tuning de Aprendizaje por Transferencia. Con la configuración adecuada, este modelo presenta bastante equilibrio en la precisión para clasificar ambas clases, aunque es más sensible con los tumores malignos. De esta manera se plantea implementar dichos modelos en un sistema que ayude a los médicos para el diagnóstico, aunque tienen una mejora potencial si se cuenta con una mayor capacidad computacional y un conjunto de datos más grande.

Palabras clave: Cáncer de mama, CNN, Machine Learning, Diagnóstico, Fine Tuning, Imágenes ultrasonidos

Abstract

Breast cancer is one of the diseases with the greatest impact on the female population worldwide, and early detection is essential to reducing its risk. To this end, this work aims to create a diagnostic support system based on numerical characteristics and ultrasound images of tumours. The goal is to classify tumors as benign or malignant by using machine learning models for numerical characteristics and convolutional neural networks (CNNs) for ultrasound images. In the first case, the Logistic Regression, Random Forest, and Gradient Boosting models are compared using different evaluation metrics. Among them, the Logistic Regression model with the best performance, considering the limited amount of data available, is Logistic Regression because it presents the greatest accuracy and generalization in new data. For the second case, due to the lack of computational capacity, the pre-trained DenseNet121 model is used using the Partial Fine Tuning Transfer Learning technique. With proper configuration, this model presents a fairly balanced accuracy for classifying both classes, although it is more sensitive to malignant tumors. Thus, the proposal is to implement these models in a system to assist physicians in diagnosis, although they have potential for improvement with greater computational capacity and a larger data set.

Key words: Breast cancer, CNN, Machine learning, Diagnosis, Fine tuning, Ultrasound images

Agradecimientos

A mis tutores María Jesús, Julio Sandubete y Ana Lazcano por su apoyo y guía durante la realización del TFG, que siempre han estado ahí cuando los he necesitado.

A mi familia y amigos, que me han acompañado y apoyado a lo largo de la carrera.

Y a todos los profesores que he tenido, que me han enseñado muchas cosas útiles para mi crecimiento profesional y personal.

ÍNDICE

Índice de Figuras

Índice de Tablas

OBJETIVOS

Objetivo General

El objetivo general de este proyecto consiste en elaborar un sistema que sea capaz de detectar de manera automática y rápida si un paciente tiene cáncer de mama, ofreciendo así un apoyo para los médicos. Para ello se hará uso de modelos de aprendizaje automático de predicción de clases.

Objetivos Específicos

Para alcanzar el objetivo general se han planteado cuatro objetivos específicos con el objetivo de guiar el proyecto. Estos son:

- Analizar las relaciones de las características de los tumores con el diagnóstico.
- Clasificar los tumores en función de las características numéricas.
- Identificar el modelo que tenga el mejor rendimiento para clasificar los tumores en base a características numéricas.
- Clasificar los tumores en base a imágenes ultrasonidos

Para resolver el primer objetivo específico se realizará un análisis exploratorio de las características de los tumores y se estudiará la relación existente con el tipo de tumor, benigno o maligno. Este análisis es fundamental para comprender el impacto de cada una de las características sobre el diagnóstico.

Para el segundo objetivo se elaborarán modelos de clasificación como Regresión Logística, Random Forest y Gradient Boosting, los cuales se utilizarán para clasificar los tumores en benignos o malignos en base a sus características.

Para el tercer objetivo se realizará una evaluación completa de los tres modelos, y posteriormente se realizará una comparativa entre estos para la elección del que ofrezca el mejor rendimiento y así utilizarlo en el sistema de apoyo al diagnóstico del cáncer de mama.

Finalmente, para llevar a cabo el cuarto y último objetivo se elaborará un modelo de Redes Neuronales Convolucionales (CNN), el cual analizará y clasificará imágenes de ultrasonidos en las clases:

- **Benigno**: masa mamaria con tumor no cancerígeno
- **Maligno**: masa mamaria con tumor cancerígeno

Gracias a la definición de estos objetivos específicos se asegura el cumplimiento del objetivo principal de este proyecto garantizando que los modelos son capaces de realizar un diagnóstico correcto y así ofrecer un sistema de detección temprana que sirva de apoyo a los médicos.

CUERPO DE LA MEMORIA

Introducción

El cáncer de mama surge por un crecimiento anómalo y descontrolado de células en las glándulas mamarias y es uno de los tipos de cáncer más comunes a nivel mundial, siendo especialmente frecuente en mujeres, ya que un 99% de los diagnósticos se asocia a este género (Novartis, s.f.).

Se trata del tercer tipo más frecuente en la sociedad española, siendo solo superado por el cáncer de próstata y el de colón.

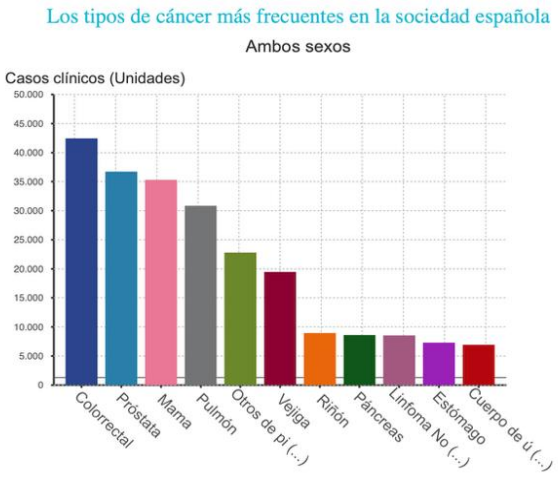

Ilustración 1. Tipos de cáncer más frecuentes en la sociedad española. Fuente: (Asociación Española Contra el Cáncer, 2024)

Se han diagnosticado hasta 35.875 casos de cáncer de mama en el año 2024 y se puede observar como la incidencia va en aumento en los últimos 7 años.

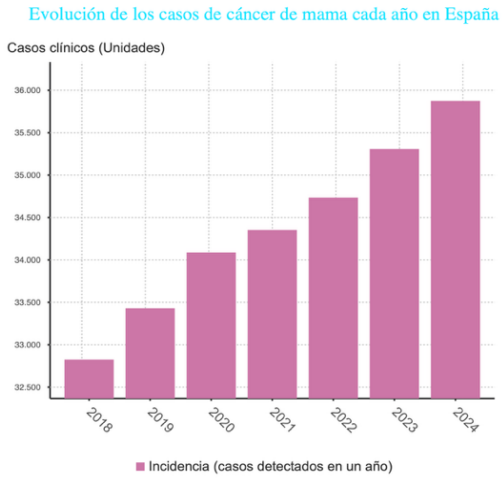

Ilustración 2. Evolución de los casos de cáncer de mama cada año en España. Fuente: (Asociación Española Contra el Cáncer, 2024)

Además, los casos detectados aumentan conforme a la edad y esto se puede observar claramente en el gráfico que se muestra a continuación, donde la cantidad de casos por cada 100.000 habitantes aumenta a partir de los 40 años y es a partir de los 60 años en adelante cuando la concentración de casos es mucho mayor.

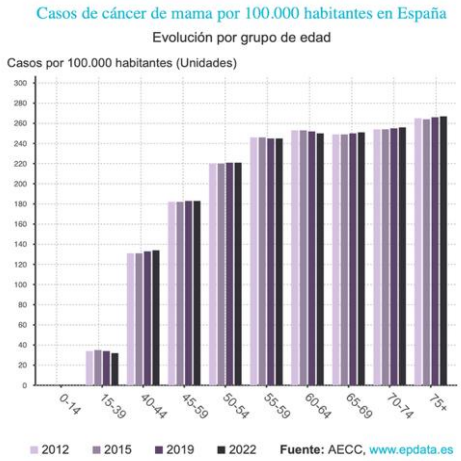

Ilustración 3. Casos de cáncer de mama por 100mil habitantes en España. Fuente: (Asociación Española Contra el Cáncer, 2024)

El cáncer de mama se origina generalmente en las células de los lobulillos (glándulas productoras de leche) o en los conductos, que son las vías que transportan la leche hasta el pezón.

Cuando el cáncer de mama se denomina invasivo (o infiltrante), significa que se ha extendido a los tejidos mamarios cercanos. Los dos tipos más frecuentes son:

- El **carcinoma ductal invasivo (CDI):** este comienza en los conductos lácteos y es el más frecuente ya que alrededor del 80% de todos los tipos de cáncer son de este tipo.
- El **carcinoma lobular invasivo (CLI):** este comienza en los lobulillos y es el segundo tipo de cáncer de mama más frecuente, con alrededor de un 10% de los casos.

Cuando el cáncer de mama se clasifica como no invasivo significa que no se ha extendido más allá del tejido mamario donde se originó. Los dos tipos más frecuentes son:

- El **carcinoma ductal in situ (CDIS):** este tipo de cáncer no se extiende más allá del punto de origen que son los conductos lácteos. Este no es potencialmente mortal, pero se considera un precedente del cáncer de mama invasivo y aumenta el riesgo de padecer dicha enfermedad más adelante.
- El **carcinoma lobular in situ (CLIS):** este tipo de cáncer no se extiende más allá de los lobulillos, que son las glándulas productoras de leche y se trata de una afección benigna.

Adicionalmente, existen dos tipos de tumores menos comunes que son:

- **Tumores phyllodes cancerosos de la mama:** estos son poco comunes y representan menos del 1% de todos los tumores mamarios. Además, la mayoría de estos son benignos, aunque cerca del 25% son cancerosos.
- **Tumores fungiformes de la mama:** estos crecen a través de la piel de la mama y solo entre el 2% y el 5% de los casos de cáncer de mama localmente avanzados se convierten en lesiones de mama fungiformes (DePolo, 2025).

La mayor parte de los tumores de mama tienen un origen en el que interviene el azar, pero entre el 10 y el 15 por ciento de ellos casos tienen un origen hereditario. Sin embargo, a pesar del factor de aleatoriedad existente, hay determinados factores que

pueden hacer que la probabilidad de tener la enfermedad sea mayor y estos son los denominados factores de riesgo.

Tener un factor de riesgo no es una condición suficiente para padecer la enfermedad y de hecho en el 50% de los casos de cáncer de mama no es reconocible ningún factor de riesgo.

Se han identificado una serie de factores de riesgo relacionados con el cáncer de mama y estos se han clasificado en dos grupos:

- Factores de riesgo **modificables**: estos son factores que pueden ser alterables con el objetivo de contribuir a la prevención del desarrollo del cáncer y son los factores hormonales exógenos, la vida reproductiva, los factores dietéticos, el ejercicio físico, la obesidad, el alcohol, el tabaco, la deficiencia de la vitamina d y las radiaciones ionizantes.
- Factores de riesgo **no modificables**: estos incluyen los aspectos no modificables de una persona como son el sexo, la raza, la edad, los antecedentes familiares, las mutaciones genéticas, factores hormonales y la densidad del tejido mamario.

Como ya se ha mencionado anteriormente, existen situaciones especiales relacionadas con el cáncer de mama hereditario en el que esta enfermedad aparece como fruto de una alteración genética en genes de alta penetrancia y que se transmite de padres a hijos en la familia. Las principales mutaciones genéticas identificadas son:

- Las más conocidas por la alta penetrancia se producen en los genes **BRCA1, BRCA2 y PALB2** que conllevan un riesgo muy alto de desarrollar cancer.
- Otros genes implicados en cáncer hereditario de menor penetrancia son **ATM, CHEK2, RAD51C, PTEN, P53,** etc (Asociación española contra el cáncer, 2023).

A pesar de esto, es necesario que las mujeres se realicen exámenes cada 1 o 2 años para conocer si tienen o no el cáncer, sobre todo las que presentan una mayor probabilidad de riesgo o aquellas que superan los 40 años. Entre estos exámenes se encuentran las mamografías, ultrasonidos, resonancias, biopsias y tomografías. Cada prueba corresponde a una fase distinta del cáncer; por ejemplo, la mamografía, los ultrasonidos y las biopsias sirven para identificar si existe un bulto en la mama y las características de este y, en cambio, las tomografías sirven para identificar si se ha extendido el tumor.

Además, el tratamiento dependerá del estado en el que se encuentre el cáncer. Este puede presentarse en 5 estados o estadios. Los estadios 0, I, II o III son los más iniciales y su tratamiento se basa en erradicar el cáncer e impedir que reaparezca, mientras que, si se encuentra en un estadio IV, se basa en intentar disminuir los síntomas y alargar el tiempo de vida, ya que en este estado se considera prácticamente incurable. Incluso si superan el cáncer, los pacientes deberán seguir tomando medicamentos y haciéndose pruebas para vigilar su reaparición. Para conocer el estadio en el que se encuentra también es importante estar pendiente de los síntomas, ya que en las fases iniciales no se suele sufrir ninguno, pero según va creciendo el tumor pueden aparecer bultos, líquidos o cambios notables en la mama o en el pezón; y cuando el tumor se encuentra en un estado muy avanzado se pueden sufrir dolores, dificultad respiratoria, hinchazón y pérdida de peso (MedilinePlus, 2024).

Motivación

Uno de los factores clave para aumentar la tasa de curación del cáncer de mama es la detección precoz. Siempre que este se encuentre en una fase o estadio inicial, las probabilidades de superarlo son significativamente más altas y, en países como España, pueden acercarse al 100% (Asociación Española Contra el Cáncer, 2023).

Aunque el cáncer de mama sigue siendo una de las principales causas de mortalidad entre las mujeres, los avances y las investigaciones han contribuido a la reducción progresiva de la tasa de mortalidad, ya que permiten a los profesionales sanitarios realizar un diagnóstico más temprano (Mayo Clinic, 2025). Sin embargo, en estadios avanzados, especialmente cuando hay presencia de metástasis, es muy complicado de superar. Por ello, identificar la enfermedad en fases tempranas es esencial (Harbeck, y otros, 2019).

En este contexto, el presente proyecto surge con el objetivo de detectar el cáncer de mama en etapas tempranas, beneficiando tanto a pacientes como al personal sanitario. Esta herramienta permite optimizar el trabajo y que los médicos se centren en la planificación y ejecución del tratamiento.

Estado del arte

La detección temprana de tumores es fundamental para salvar vidas, y es por eso que se han realizado muchos estudios con este objetivo basados en técnicas de Machine Learning y Redes Neuronales.

Entre los modelos utilizados se encuentran el clasificador Naive Bayes (NB) y el vecino más cercano (KNN), los cuales presentan una precisión de 96,19% y 97, 51% respectivamente (Amrane, Oukid, Gagaoua, & Ensarİ, 2018)

En cuanto al clasificador Naive Bayes, también se ha experimentado con distintos modelos de este tipo como el Naive Bayes aumentado a árbol (TAN), Clasificador Bayesiano K-Independiente (KDB) y Naive Bayes aumentado a bosque (FAN). Se concluyó que estos modelos son bastantes adecuados, ya que permiten hacer una buena clasificación basándose en las dependencias entre las distintas características del tumor, puesto que en los resultados presentaron buenas métricas como un 81% en exactitud, un 65% en sensibilidad y un 89% en especificidad (Gabriel, López, & Barbosa, 2013)

De la misma manera, también se han realizado estudios comparativos, como el que se pretende en este proyecto, de distintos modelos de clasificación tanto de aprendizaje supervisado como no supervisado. Entre estos modelos se encontraban la Regresión Logística, KNN, K-Means, Random Forest, Support Vector Machine (SVM), Análisis de Discriminantes Lineales (LDA), Gaussian Naive Bayes y Multilayer Perceptron (MLP). En esta comparación, se encontró que el modelo que ofrece mejor rendimiento en el entorno general era el MLP, mientras que dentro de los supervisados era el SVM y de los no supervisados el K-Means (Gabriel Mauricio Martínez-Toro, 2018).

Otros estudios se centraron más en los modelos de aprendizaje supervisado, como Regresión Logística, Árboles de Decisión, Gaussian Naive Bayes, KNN, Random Forest y SVM. De estos se llegó a la conclusión de que el modelo que mayor precisión ofrecía eran los Árboles de Decisión con un 100%, mientras que el peor fue el KNN con un 94%. De todas maneras, se observa que estos modelos ofrecieron un gran rendimiento para clasificar los tumores de cáncer de mama (Jorge Armando Millán Gomez, 2020).

Algunos de estos modelos también fueron entrenados siguiendo las categorías establecidas en el BI-RADS (Breast Imaging Reporting and Data System), que se basan en la densidad del tejido mamario, ya que según la Organización Mundial de Salud (OMS) las mujeres con tejidos más densos tienen mayor riesgo de desarrollar cáncer de mama. Entre

los modelos que se entrenaron se encuentran de nuevo SVM, Random Forest, Regresión Logística y KNN, además de AdaBoost. Tras una comparación con distintos clasificadores, se concluyó que el que ofrecía mayor rendimiento era el SVM, aunque la mayoría también ofrecía buenos resultados con métricas superiores al 80% (Dieazago Alejandro Arturo Angulo, 2024)

Estos modelos has sido probados para el mismo propósito en numerosas ocasiones, por lo que se volvieron a probar el Random Forest y SVM, a los que se les añadió Redes Neuronales. Sin embargo, en este caso, el SVM volvió el modelo que clasificaba con mayor precisión si el tumor era benigno o maligno, con un 99%. Pero, a pesar de que el SVM fue el modelo más preciso, el resto también tuvieron un rendimiento bastante alto respecto a la exactitud de la clasificación, por lo que cualquiera de estos se podría aplicar y técnicas de hospitales y laboratorios para detectar el cáncer de mama (Collazo, 2020).

También se han estudiado otros algoritmos de clasificación con el objetivo de reducir el número de biopsias de mama para facilitar el diagnóstico. Entre estos estos algoritmos se encuentran MLP, función de base radial (RBF) y redes neuronales probabilísticas (PNN). Estos también presentaron muy buenos resultados, destacando las PNN con un 100% y 97,66% en las fases de entrenamiento y prueba, respectivamente. También la MLP tuvo buenos resultados con un 97,80% y 96, 34%, aunque la RBF fue mejor durante el entrenamiento, pero peor en la validación (Azar & El-Said, 2013).

Uno de los algoritmos más explotados para este caso también ha sido son las redes neuronales, ya que existen muchos tipos que pueden ser de gran utilidad, como las PNN, mencionadas anteriormente, las redes neuronales artificiales (ANN) o las redes neuronales convolucionales (CNN). Respecto a las ANN, han existido muchos estudios, pero no se han conseguido implementar en la clínica debido a una cierta desconfianza de los médicos y de los pacientes a los diagnósticos de una ‘máquina’. Además, aunque este modelo tuvo éxito en algunos casos y en otros nos, en los que lo tuvo era necesario un alto coste computacional, por lo que no era óptimo (Abbass, 2002).

En cuanto al CNN, se encontró que mejoraba modelos como el Resnet 50. Aunque en las primeras fases pueda sobre ajustarse, es fácil optimizarlo al mejorar parámetros, llegando a tener hasta un 88% de tasa de reconocimiento de las distintas anomalías. Gracias a esto se concluyó que es una técnica muy útil para la detección temprana, pero que se podría mejorar mucho más con grandes conjuntos de datos de imágenes de distintas

condiciones o incluso combinándolo con técnicas de inteligencia artificial (Khan, y otros, 2021).

Otros trabajos también hicieron una comparativa entre distintos modelos preentrenados de CNN como EfficientNetB0, ResNet50, ResNet152, EfficientNetV2 y ConvNeXt. Entre estos modelos se encontró que el más eficiente era EfficientNetB0 con métricas del 77% en sensibilidad y 97.03 en especificidad. Sin embargo, se mantuvo a Resnet50 como una buena alternativa con menos recursos, ya que con una buena cantidad de datos presentaba un AUC-ROC de 90% (Anthony Esteban Aldaz Noble, 2025).

También se ha encontrado que otro modelos preentrenados como DenseNet121 presentan una gran eficiencia en el diagnóstico de asimetrías en mamografías. Se ha concluido esto, ya que ha presentado métricas con valores de 87% en precisión y 90% en especificidad tras entrenarlo con imágenes de 460 pacientes del Hospital Shenzhen de Pekin (Tingting Liao, 2023).

Sin embargo, en otras comparativas se ha encontrado que modelos como ResNet50 y SheffleNet presentan una precisión ligeramente más alta que DenseNet121 al aplicar la técnica de preprocesamiento CLAHE en las imágenes. Sin embargo, sin aplicar la técnica, los tres modelos presentaban la misma precisión del 75% A pesar de esto, los tres modelos presentaron altos valores de precisión con y sin la técnica (Pedro Moises de Sousa, 2024).

Otros estudios han hecho comparaciones entre más modelos de CNN para realizar diagnósticos sobre el riesgo del cáncer de mama a través de mamografías. Entre estos modelos se encuentran MobileNetV3 Small, MobileNetV3 Large, MobileNet, ResNet50, ResNet152, VGG16, VGG19, DenseNet121, InceptionV3 y AlexNet12. Se concluyó que los más eficientes fueron Renste50 y EfficientB7, pero se determinó que tienen sus fortalezas y la elección de cuál utilizar depende del caso y del conjunto de datos (Chávez, 2023).

Para confirmar la eficacia de los modelos de CNN preentreneados y personalizados mediante aprendizaje por transferencia, se han realizado otros estudios con modelos como DenseNet entrenado con características de la base de datos de imágenes INbreast. En este caso, este modelo presentó métricas de 99%, para exactitud y especificidad, 98% de sensibilidad y 96% de precisión. Estos valores tan altos presentan que el uso de modelos preentrenados es una técnica muy buena y precisa para realizar diagnósticos con mamografías (Abeer Saber, 2023).

Además, se hicieron otros estudios para comparar la capacidad de diagnóstico de un médico experimentado contra un sistema automático con modelos de CNN preentrenados como GG16, VGG19, RESNET50 y uno desarrollado desde cero. En este caso se comprobó que un sistema que utilizaba la estructura de VGG19 conseguía tasas de acierto superiores al 90%, mientras que la capacidad del médico alcanzaba un 79.97% (Harold Agudelo Gaviria, 2021)

En conclusión, este es un tema para el que se han realizado muchos estudios y se ha experimentado con diversos modelos; pero en todos estos, se ha determinado que los óptimos son los modelos de clasificación y sobre todo las redes neuronales, teniendo en cuenta que cuando se trata con imágenes lo mejor son modelos preentrenados de CNN. En este proyecto se presenta un enfoque similar ya que también se intentará optimizar el uso de los modelos de CNN para conseguir detectar lo más rápido posible si un tumor es benigno o maligno. De la misma manera, también se utilizarán otros modelos de clasificación con el mismo objetivo cuando se cuenta con las características numéricas de los tumores.

Marco teórico

El cáncer de mama es uno de los más comunes en la sociedad, especialmente en las mujeres y la detección temprana es fundamental para poder llevar a cabo los tratamientos necesarios para salvar la vida de los pacientes. Por ello, la realización de pruebas es muy importante y el posterior análisis de los resultados por parte del personal sanitario es esencial. Para esto último, es fundamental que el personal médico cuente con los sistemas y los recursos necesarios para ofrecer el mejor diagnóstico a los pacientes.

Para ofrecer apoyo a los médicos a la hora del análisis de las diferentes pruebas realizadas a pacientes, en este proyecto se elaborarán modelos de aprendizaje automático y de redes neuronales con el objetivo de identificar la existencia de tumores y si estos son cancerosos o no. Estos modelos son: Regresión Logística, Random Forest, Gradient Boosting y CNN.

A continuación, se va a explicar por qué se han escogido estos modelos y los fundamentos de cada uno.

Regresión Logística

Se trata de un algoritmo de machine learning supervisado, formulado por David Cox en 1958, quien analizó las dependencias que hay entre los valores independientes correspondientes a un ensayo y las probabilidades de que la respuesta de este sea 0 o 1 (Cox, 1958). Este modelo no es un simple algoritmo que clasifica, sino que se trata de un modelo probabilístico que calcula la probabilidad condicional de una clase dada una combinación lineal de predictores.

Se utiliza para resolver problemas de clasificación binaria mediante la predicción de la probabilidad de pertenencia a una clase concreta. Esto se logra mediante la relación entre las variables dependientes y la variable objetivo a través de la función logística.

La función logística o sigmoidea busca calcular la probabilidad de las instancias, es decir, obtener valores entre 0 y 1, pero con el objetivo de clasificarlas solamente como 1, cuando su probabilidad es mayor que el umbral establecido, o 0, cuando su probabilidad es menor de dicho umbral. De esta manera, la curva de la función sigmoidea o función logística tiene forma de "S", la cual se presenta como $d(z) = \frac{1}{1+e^{-z}}$.

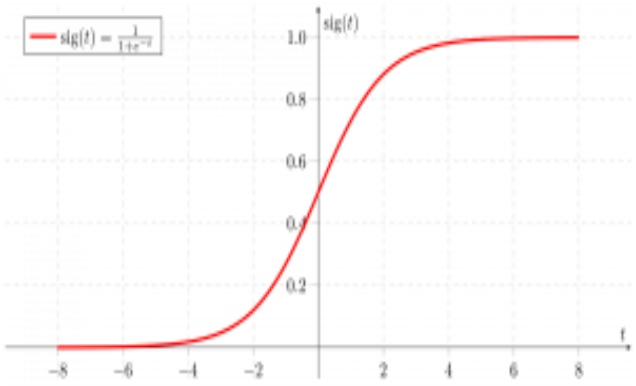

Ilustración 4. Función Sigmoidea. Fuente: (Geeks for Geeks, 2024)

Este modelo representa la relación lineal entre la variable independiente y el logaritmo de la probabilidad de la variable dependiente, lo que significa que los predictores afectan al logaritmo de las probabilidades de forma lineal (Geeks for Geeks, 2024)

Se ha seleccionado por su simplicidad y eficacia en tareas de clasificación binarias como la predicción de tumores benignos o malignos. También es un modelo muy ligero en comparación con otros por lo que se entrena más rápido y requiere menos recursos computacionales. Además, es un modelo que funciona bien cuando existen altas correlaciones entre variables y por lo tanto cuando existen relaciones lineales entre estas.

Random Forest

Este algoritmo de machine learning, registrado por Leo Breiman y Adele Cutler, consiste en combinar árboles de decisión para alcanzar un resultado optimizado. Se trata de un modelo útil tanto para problemas de regresión como de clasificación (IBM, s.f.).

Los árboles de decisión están compuestos por nodos de hojas y sus ramas. Los nodos serían las preguntas a las que responde el modelo para dividir los datos en ramas que corresponderán a las distintas respuestas (GeeksforGeeks, 2025). El objetivo de este algoritmo es encontrar la mejor división del subconjunto de datos para seguir la ruta que lleve al resultado esperado. Aunque se trata de un algoritmo supervisado es muy propenso a problemas como el sobreajuste, por lo que un conjunto de árboles en un algoritmo de bosque aleatorio es mejor, ya que busca el resultado óptimo entre todos los árboles (GeeksforGeeks, 2025).

Los métodos de aprendizaje por conjuntos se componen por varios clasificadores para agregar sus predicciones y así encontrar el mejor resultado. Los más conocidos son el bagging y el boosting. En 1996 Leo Breiman desarrolló el método de bagging, el cual se basa en seleccionar varias muestras aleatorias en un conjunto de datos con reemplazo, es decir que los datos pueden aparecer más de una vez, para entrenar los modelos de forma independiente de manera que el promedio o la mayoría de las predicciones muestren una estimación más precisa. Esto permite reducir la varianza en un conjunto de datos ruidoso (GeeksforGeeks, 2025).

El Random Forest es una extensión del bagging ya que lo utiliza, así como la aleatoriedad de características, para crear un bosque no correlacionado de árboles de decisión. Esta es la diferencia entre los árboles de decisión y el random forest, ya que, mientras que los árboles tienen en cuenta todas las posibles divisiones de las características, los bosques solo seleccionan un subconjunto aleatorio.

Este modelo está compuesto por un conjunto de árboles aleatorios definido y cada árbol por un subconjunto de datos aleatorio extraído de un conjunto de entrenamiento con reemplazo, reservando un tercio de los datos como prueba. Esto hará que en los árboles predomine la variabilidad reduciendo el sobreajuste y mejorando el rendimiento del modelo, ya que aprenderá sobre todas las relaciones posibles. Al ser un modelo útil tanto para tareas de regresión como de clasificación, la determinación de la predicción variará,

siendo el promedio de los resultados de los árboles para la regresión y la variable categórica más frecuente para la clasificación (IBM, s.f.).

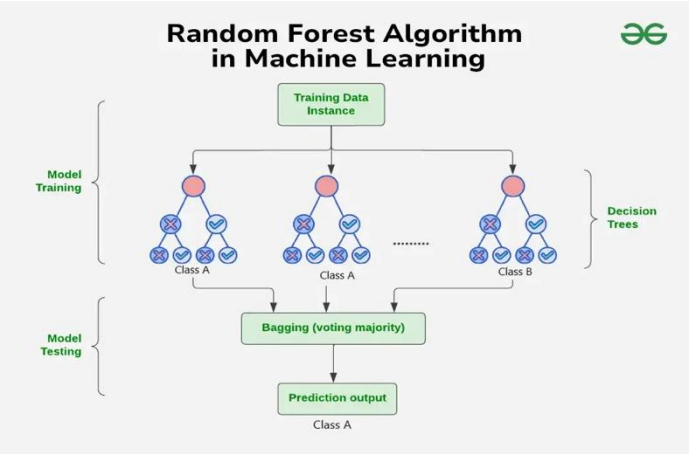

Ilustración 5. Ejemplo Random Forest. Fuente: (GeeksforGeeks, 2025)

Este modelo ha sido seleccionado por la gran capacidad que tiene para manejar relaciones no lineales entre los datos de los distintos tumores. Esto se debe a que, como se ha explicado anteriormente, es un modelo que está basado en árboles de decisión y por lo tanto no asume relaciones lineales entre las variables. Por otro lado, como ya se ha comentado anteriormente, un solo árbol tiende al sobreajuste, pero al hacer uso de una gran cantidad de árboles y siendo cada uno de estos entrenado con una muestra distinta, la generalización mejora considerablemente.

Gradient Boosting

Gradient Boosting es un modelo que consiste en crear varios predictores en secuencia, lo que quiere decir que cada modelo aprende los errores del anterior para encontrar el mejor resultado (Barajas, 2024).

Fue un modelo propuesto por Friedman en 1999, quien en su estudio (Friedman, 1999) consideraba que la optimización numérica en la estimación de funciones no se encontraba en los parámetros sino en las propias funciones. Además, estableció que existía una conexión entre las sumas secuenciales de funciones y la minimización de los errores.

Al igual que el Random Forest, se trata de un algoritmo de machine learning supervisado por conjuntos que combina múltiples modelos débiles para crear uno final más robusto. Pero, en lugar de árboles, este modelo se basa en entrenar secuencialmente

modelos débiles añadiendo más pesos a aquellos que presenten mayor tasa de error, que es la diferencia entre los valores reales y las predicciones. Esto sirve para calcular el gradiente que se utiliza para encontrar la dirección del ajuste de parámetros del siguiente modelo y así minimizar la función de pérdida.

Se diferencia de los modelos de bagging, como el random forest ya que, en lugar de optimizar la predicción a través de entrenamientos iterativos, lo hace a través de entrenamientos secuenciales. Además, ambos métodos se diferencian de las redes neuronales, ya que buscan la mejor predicción utilizando modelos débiles en lugar de un modelo muy complejo (Hoss Belyadi, 2021).

Este modelo se ha elegido ya que, al igual que el Random Forest y la Regresión Logística, es muy útil para problemas de clasificación binaria, además de que permite trabajar con relaciones complejas y optimiza los resultados gracias al entrenamiento secuencial. Aunque es más complejo de desarrollar que el Random Forest, debido a su sensibilidad a los parámetros, suele ser mucho más preciso y eficiente en términos de tiempo y recursos. También es un modelo eficaz a la hora de manejar el desbalanceo de clases, situación que en problemas médicos sucede mucho, porque se centra más en los casos mal predichos. Esto permitirá clasificar los tumores en benignos o malignos de una forma optimizada mediante las relaciones entre sus características.

Redes Neuronales Convolucionales (CNN)

El modelo CNN, creado por Yann LeCun en 1989, es un algoritmo de aprendizaje profundo especializado en tareas de procesamiento de imágenes (Aprende Machine Learning, 2018).

Consiste en un algoritmo con una arquitectura que le permite ser muy útil para tareas de clasificación de imágenes. Esta arquitectura consiste en:

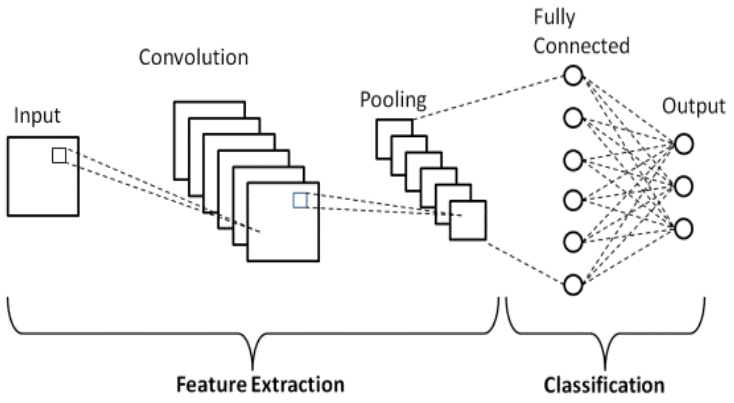

Ilustración 6. Capas CNN. Fuente: (Melo, 2023)

- Capas de entrada, que es la imagen o secuencia de imágenes sin procesar que se utilizan como entrada.
- Capas convolucionales donde se detectan las características de las imágenes aplicando operaciones convolucionales mediante filtros o núcleos, los cuales se definen como Kernel. Esta convolución consiste en la transformación de una función en base a otra por el producto escalar, es decir, que dichas operaciones convolucionales se refieren al producto escalar entre el Kernel y las ventanas de la imagen del mismo tamaño para transformar dicha imagen en las características que servirán de entrada para la red neuronal. Esta operación se repetirá para todas las ventanas hasta completar la imagen. El Kernel se trata de una matriz constante de pesos predefinida, generalmente de 2x2, 3x3 o 5x5, dependiendo del objetivo que se tenga, como desenfocar la imagen, enfocarla, detectar formas, etcétera.
- Capa de activación donde se le aplica la función de activación a las salidas de la capa anterior para agregar no linealidad a la red y de esta forma aprender las relaciones más complejas en los datos. Esto también puede presentarse dentro de la capa de convolución.
- Capas de agrupación o pooling donde se reduce el tamaño de las dimensiones espaciales de la entrada reduciendo la complejidad y la cantidad de parámetros en la red. Existen dos tipos de agrupación muy comunes como la agrupación máxima y la agrupación promedio, las cuales pueden afectar a la entrada por porciones de un tamaño determinado o directamente a toda la entrada, lo que se denominaría como global.

- Capa de aplanamiento donde las características resultantes de la capa anterior se aplanan en un vector unidimensional
- Capas totalmente conectadas donde se conectan las neuronas de cada capa con la siguiente y se realizan las tareas de predicción o clasificación de la imagen basándose en las características de alto nivel aprendidas por las capas anteriores.
- Capa droput donde se realiza una regularización en la que algunas neuronas elegidas aleatoriamente son ignoradas de forma temporal durante el entrenamiento para evitar el sobreajuste.
- Capas de salida donde se ejecuta la función de clasificación para obtener los outputs finales. Entre estas funciones se puede encontrar la Regresión Logística o Softmax, despendiendo si la clasificación es binaria o multiclase, respectivamente.

Esta arquitectura le permite asemejarse al proceso visual del cerebro humano, por lo que es útil para capturar patrones jerárquicos y dependencias dentro de las imágenes (Geeks for Geeks, 2024)

Además, para desarrollar un modelo CNN se puede hacer desde cero o mediante aprendizaje por transferencia. Esto último es una técnica que consiste en utilizar modelos preentrenados con grandes cantidades de datos para un propósito concreto y personalizarlo para otro propósito. Existen varios tipos de técnicas, llamadas Fine-Tuning, para personalizar estos modelos durante el entrenamiento:

- Full Fine Tuning que consiste en personalizar toda la red neuronal
- Partial Fine Tuning que consiste en personalizar solo un conjunto de capas, que generalmente son las últimas, ya que son donde se realiza la tarea de clasificación.
- Feature Extraction Fine Tuning que consiste en reemplazar únicamente la capa de salida, ya que por capas específicas de la tarea de clasificación que se desea realizar.m
- Discriminative Fine Tuning que consiste en aplicar tasas de aprendizaje progresivas, es decir tasas más bajas en las primeras capas y tasas más altas en las últimas capas.

Este tipo de técnica es útil cuando no se tiene suficiente capacidad computacional o recursos para desarrollar y entrenar un modelos desde cero (Bergmann, 2024).

Por esto se ha escogido este modelo, ya que también se van a utilizar imágenes de ultrasonidos para identificar si masas mamarias contienen tumores o no y, si es el caso, si este es maligno o benigno. Este modelo permitirá procesar dichas imágenes mediante su estructura de capas, desde la entrada de las imágenes hasta la predicción de las clases

pasando por el procesamiento mediante el Kernel y las operaciones convolucionales, la identificación de las relaciones mediante la función de activación, la reducción de su complejidad y la reducción de los modelos de clasificación. Además, permite el uso de modelos muy potentes ya preentrenados con una simple personalización. Gracias a esto se podrán clasificar la imágenes mediante las relaciones que presenten sus formas, colores, etcétera.

TRABAJO TÉCNICO

Herramientas empleadas

Para el desarrollo de este trabajo se ha hecho uso del lenguaje de programación Python ya que ofrece una gran cantidad de librerías para realizar tanto el análisis de datos como la elaboración de los modelos de machine learning y de redes neuronales. Entre estas librerías se encuentran Pandas para la lectura y el tratamiento de los datos, Matplotlib para crear los gáficos, Sklearn para el desarrollo de los modelos y sus métricas de evaluación y Tensorflow y Keras para el preprocesamiento de imágenes y desarrollo del modelo CNN. También, para poder desarrollar el CNN, se ha utilizado Google Colab, ya que posee GPUs muy potentes para tratar con imágenes, como la A100.

Ingeniería del dato

Se cuentan con dos tipos de datos, unos numéricos como características de tumores y otros de imágenes ultrasonidos de masas mamarias.

Características de tumores

Origen de los datos

En cuanto a las características de los tumores se tratan de 569 casos, proporcionados por el Dr. Wolberg del Hospital Universitario de Wisconsin. Se trata de muestras recogidas de imágenes digitalizadas de pruebas por aspiración de aguja fina (PAAF) de masas mamarias.

Fuente: https://archive.ics.uci.edu/datasets?search=Breast%20Cancer%20Wisconsin%20(Original)

Características de los datos

Estos datos cuentan con 569 observaciones y 32 variables, las cuales, realmente, son 10 características correspondientes a cada muestra recogida por el PAAF (se recogieron muestras de 3 células) más el ID y el Diagnóstico del caso. En este caso, se ha decidido realizar la media de las tres muestras, para mantener únicamente 10 variables y así evitar redundancia de información que pueda afectar a los modelos.

Tabla 1. Descripción de las variables numéricas

Variable	Descripción	Tipo
ID	Número para identificar el caso	Entero
Diagnosis	M = Maligno B = Benigno	Objeto
Radius	Media de las distancias desde el centro a los puntos del perímetro	Flotante
Texture	Desviación estándar de los valores de escala de grises	Flotante
Perimeter	Perímetro de la célula	Flotante
Area	Área de la célula	Flotante
Smoothness	Variación local en longitudes de radio	Flotante
Compactness	Perimeter^2 / Area - 1.0	Flotante
Concavity	Severidad de las porciones cóncavas del contorno	Flotante
Concave_points	Número de porciones cóncavas del contorno.	Flotante
Symmetry	Nivel de simetría de la célula	Flotante
Fractal_dimension	"Aproximación de la línea costera" - 1	Flotante

Elaboración Propia

Transformaciones de los datos

A estas variables se les han aplicado una serie de cambios para que tengan el formato necesario para los modelos que se van a utilizar:

- Para comprender mejor la variable ‘Diagnosis’ se han cambiado los valores por 1 = Maligno y 0 = Benigno. De esta manera se ha transformado el tipo de dato a ‘Entero’.

Estudio estadístico

Con los datos limpios y preparados, se hizo un análisis exploratorio para analizar toda la información que aportan y cómo puede utilizarse para conseguir los objetivos.

Para ello primero se hizo un estudio estadístico para obtener una visión a cerca de las tendencias de los datos.

Tabla 2. Valores estadísticos de las variables numéricas

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Diagnosis	569	0.37	0.48	0	0	0	1	1
Radius	569	10.27	2.83	5.04	8.35	9.57	11.69	22
Texture	569	15.39	3.49	7.38	12.75	15.17	17.54	28.33
Perimeter	569	67.36	19.65	31.92	54.03	62.29	77.39	152.25
Area	569	525.4	316.6	112.84	319.81	422.98	627.16	2432.4
Smoothness	569	0.08	0.01	0.05	0.07	0.08	0.08	0.12
Compactness	569	0.13	0.07	0.02	0.08	0.11	0.17	0.46
Concavity	569	0.13	0.1	0	0.05	0.11	0.18	0.62
Concave_points	569	0.06	0.03	0	0.03	0.05	0.08	0.17
Symmetry	569	0.16	0.03	0.09	0.14	0.16	0.18	0.33
Fractal_dimension	569	0.05	0.008	0.04	0.04	0.05	0.05	0.1

Elaboración Propia

En cuanto a la media y la mediana, cabe destacar que en la variable de Diagnosis la media es de 0.37, lo que quiere decir que aproximadamente un 37% de los casos son malignos, por lo que estos pueden ser datos atípicos. Para el resto de las variables se observa que las medias suelen ser altas, lo que representa una distribución asimétrica, teniendo incluso los datos con mayores valores en el último cuartil, dando lugar a outliers, lo que representa casos excepcionales de pacientes con características muy pronunciadas en sus tumores.

Respecto a la desviación estándar, presenta valores muy altos, sobre todo en el área y el perímetro, lo que representa una gran heterogeneidad en los datos y, teniendo en cuenta variables mencionadas, que puede haber tumores de distintos tamaños muy bien diferenciados permitiendo analizar si el tamaño es influyente en el diagnóstico.

Estudio descriptivo

También se han analizado las relaciones entre las distintas variables y el comportamiento de los outliers mediante gráficos descriptivos.

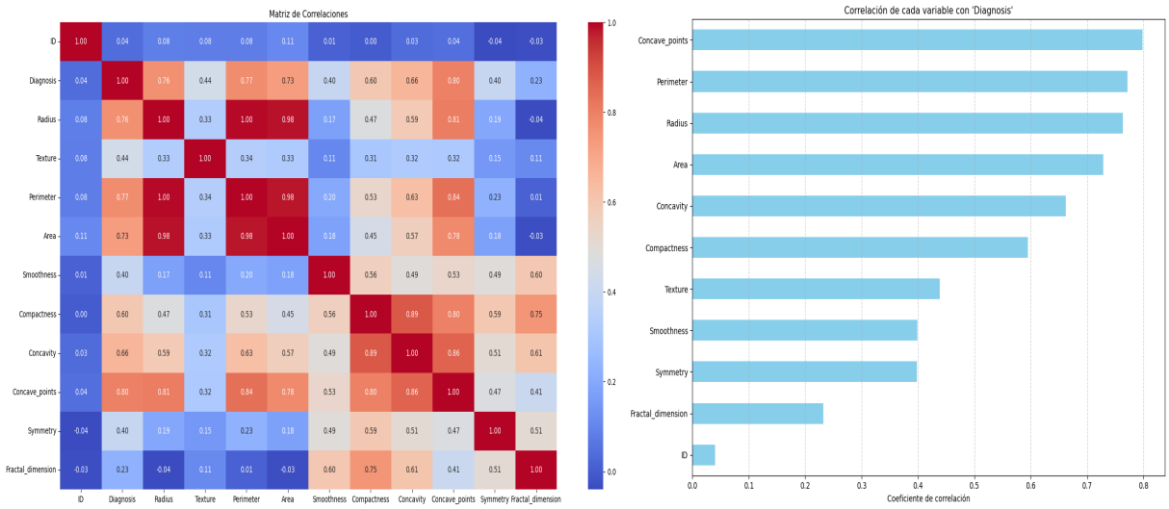

Ilustración 7. Correlaciones entre las variables numéricas y con el diagnóstico. Fuente: Elaboración propia con Python

En cuanto a las correlaciones, se observa que las relaciones con el diagnóstico son generalmente altas y directas, exceptuando la dimensión fractal, ya que es la única variable que presenta menor relación con un coeficiente 0.23, y el número de concavidades, ya que es la variable con mayor relación con un coeficiente de 0.8.

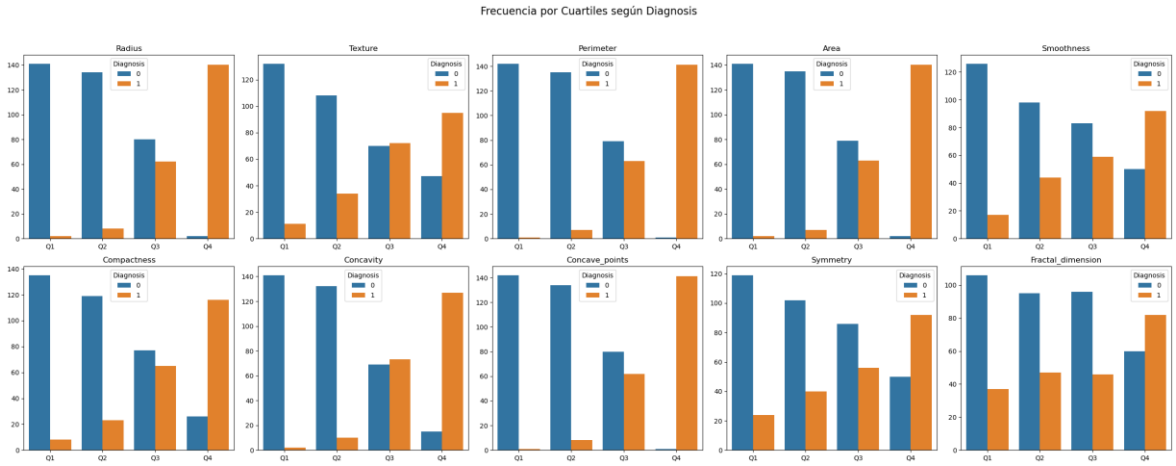

Ilustración 8. Frecuencia por cuartiles de las variables numéricas según Diagnosis. Fuente: Elaboración propia con Python

Esto se confirma en las frecuencias, ya que para todas las variables se observa una tendencia clara de crecimiento de casos malignos, y decrecimiento de casos benigno, según se avanza en los cuartiles. Este crecimiento se observa más pronunciado en el número de concavidades, ya que en el primer cuartil prácticamente no hay casos malignos, mientras que la mayoría se encuentran en el cuarto. Sin embargo, esta tendencia no es tan clara en la

dimensión fractal, ya que aunque la mayoría de los casos malignos se encuentren en el cuarto cuartil, entre el primero y el tercero prácticamente no hay diferencias, teniendo incluso un número de casos ligeramente mayor en el segundo cuartil.

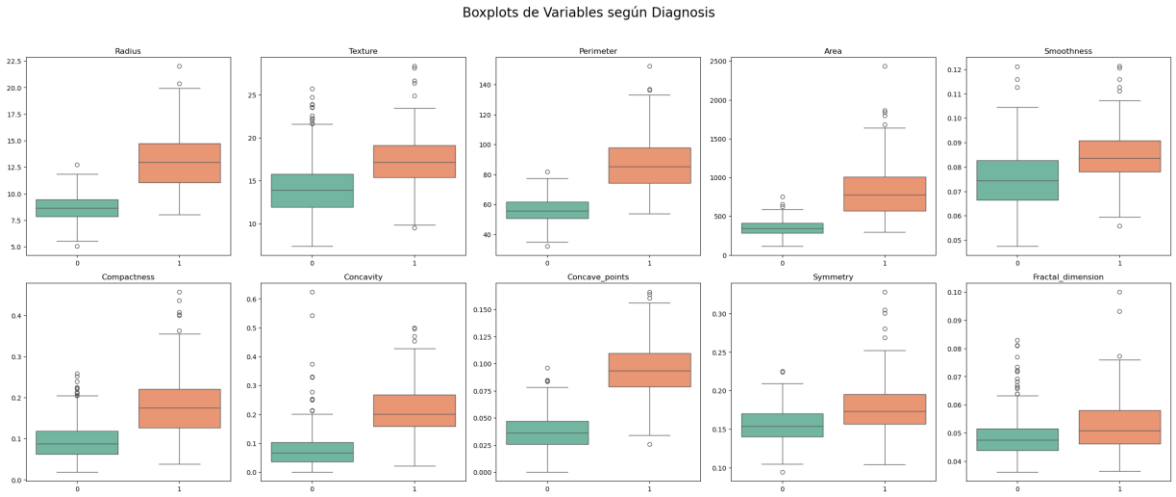

Ilustración 9. Boxplots de las variables numéricas según Diagnosis. Fuente: Elaboración propia con Python

Respecto a los outliers, se observa una clara diferenciación entre los casos malignos y benignos, ya que generalmente los malignos se asocian a los valores más altos y dispersos, puesto que las medianas de estos casos se encuentran por encima de los casos benignos e incluso muestran muchos valores atípicos en la parte alta. Sin embargo, cabe destacar variables como la dimensión fractal, donde existe casi no se diferencia entre benignos y malignos, ya que, aunque los malignos son ligeramente superiores, ambas distribuciones son muy similares, lo que quiere decir que con esta variable prácticamente no existe discriminación entre los casos. También hay que destacar la concavidad, que presenta outliers muy altos en los casos benignos, incluso superando los de los casos malignos, lo que indica que incluso dentro de los tumores benignos puede haber valores extremos.

Imágenes Ultrasonidos

Origen de los datos

Respecto a las imágenes, se trata de ultrasonidos recogidos por el profesor Aly Fahmy en 2018, de 600 pacientes mujeres de entre 25 y 75 años del Hospital Baheya de el Cairo, Egipto. Estos se dividen en:

- Normales
- Benignos
- Malignos

Cada tipo contiene su imagen ultrasonido y su máscara, que es el contorno del tumor. Fuente: https://scholar.cu.edu.eg/?q=afahmy/pages/dataset

Características de los datos

Se trata de 780 imágenes de masas mamarias, tomadas por ultrasonidos, con un tamaño medio de 500x500 pixeles en formato PNG. Entre estas se encuentran 133 imágenes de masas normales (sin tumor), 437 de masas con tumores benignos y 210 con masas de tumores malignos. Además, todas estas imágenes cuentan con sus máscaras, que es la imagen en negro con el contorno del tumor resaltado.

Preprocesamiento de imágenes

Antes de desarrollar el modelo CNN se debe preprocesar las imágenes para que contengan las características necesarias para que el modelo pueda realizar la tarea de clasificación de la mejor manera.

Primero, se decidió eliminar las máscaras, ya que la intención es que el modelo identifique el tipo de tumor con las imágenes originales, por lo que las máscaras no aportan ningún tipo de información. También, debido al desbalanceo de clases, ya que, como se ha comentado antes, hay una gran diferencia entre la cantidad de imágenes de cada clase, siendo la más pequeña la de la clase normal con solamente 133 imágenes, por lo que también se han eliminado para evitar problemas con el modelo y centrándose así en la clasificación de tumores benignos y malignos. De la misma manera, como entre estas dos clases seguía existiendo desbalanceo, ya que el número de imágenes de tumores benignos es más del doble que el de tumores malignos, se ha realizado una aumentación de las imágenes. Esto consiste en crear nuevas imágenes haciendo ligeras modificaciones a cada imagen original (Rivera, 2017). Pero hay dos tipos de aumentaciones: manual y dinámica. En este caso se han combinado los dos tipos para igualar las clases (mediante la manual) y para favorecer la variabilidad durante el entrenamiento (mediante la dinámica). Para el

primer propósito, se ha definido que por cada imagen de tumores malignos surjan 3 nuevas imágenes con las siguientes modificaciones:

- Rotación aleatoria de 30 grados
- Desplazamiento horizontal del 10% de ancho
- Desplazamiento vertical del 10% de alto
- Zoom aleatorio del 20%
- Volteo horizontal (efecto espejo)
- Relleno de pixeles vacíos con el valor más cercano

De esta manera estas nuevas imágenes se mantendrán en la memoria balanceando los datos para el entrenamiento.

También se transformaron las imágenes a un tamaño de 300x250 pixeles con formato RGB, ya que es un formato mejor aceptado en los modelos preentrenados.

Además, para facilitar el entrenamiento del modelo se han creado generadores de imágenes, que permiten cargar nuevas imágenes en lotes de forma dinámica sin guardarse en la memoria. Para ello, se han creado generadores para entrenamiento, validación y prueba. Para el generador del entrenamiento se ha aplicado la aumentación dinámica en la que se han utilizado modificaciones similares a las de la aumentación manual anterior, además de un rescaldo de los pixeles para que solo tengan valores entre 0 y 1. En cambio, para la validación y prueba, a los generadores no se les aplicó ninguna aumentación a parte del rescaldo de pixeles. También, a estos generadores se les aplicó un batch de 32, lo que quiere decir que cada paso del entrenamiento va a recibir únicamente 32 imágenes favoreciendo la eficiencia computacional, ya que el modelo irá aprendiendo por lotes en lugar de con todo el conjunto.

De esta manera los datos quedan mejor balanceados, con 734 imágenes de tumores benignos y 668 de malignos, y preparados para el desarrollo del modelo.

Análisis del dato

Características de tumores

Explicación del problema de análisis que se plantea

Se plantea un problema de clasificación binaria con el objetivo es predecir una clase dependiente de otros factores relacionados con los tumores para poder realizar un diagnóstico médico. Se buscará identificar si el tumor es benigno o maligno con la variable Diagnosis.

Para ello se ha planteado utilizar 3 modelos fundamentales de clasificación supervisada: Regresión Logística, Random Forest y Gradient Boosting. Tras el entrenamiento de estos modelos, se hará una comparación mediante distintas métricas de evaluación para evaluar cual es el mejor para cada caso.

Justificación y detalle de los modelos

A) Regresión Logística

Este modelo ha sido elegido gracias a su gran capacidad y sencillez para realizar clasificaciones binarias basándose en la dependencia lineal de las variables. Es útil para estÇe caso, ya que se busca realizar una clasificación de las clases maligno y benigno. Además, como se ha visto en el análisis descriptivo, existen altas correlaciones entre las distintas variables con el objetivo, por lo que este modelo será capaz de capturarlas para aprender cómo se comporta el objetivo en función del resto de variables y así realizar la clasificación binaria de la mejor manera posible.

Para proceder al desarrollo, primero se han elegido todas las variables como predictoras menos el ID, la Dimensión Fractal (que como se ha visto en el estudio descriptivo no ofrecía prácticamente discriminación entre las clases) y la variable objetivo, la cual ha sido Diagnosis. Tras dividir las variables, estas se han normalizado mediante la función StandardScaler ya que, al tratarse de un modelo basado en probabilidades de 0 a 1, los datos deben adecuarse a estos valores adquiriendo una media de 0 y una desviación de 1, evitando, de esta manera, que haya variables dominantes. Posteriormente se ha entrenado el modelo con un máximo de 100 iteraciones, y para ello se ha utilizado el optimizador ‘liblinear’, que es un algoritmo muy eficaz para problemas de clasificación binaria. También se aplicó una regularización C=0.1, para evitar el sobreajuste, y se

estableció que el parámetro ‘class_weight’ sea balanceado, para corregir el desequilibro que se ha observado en la clases. Además, se ha mantenido el threshold de 0.5, lo que quiere decir que, si la probabilidad calculada por el modelo es mayor a este valor, se trataría de un caso maligno.

B) Random Forest

Este modelo ha sido elegido debido a su capacidad para capturar relaciones no lineales en las variables. Esto ocurre sobre todo en estos casos clínicos, donde puede haber casos peculiares o atípicos, por lo que este modelo podrá aprender de ellos. Tampoco necesita normalización, lo que también es útil en estos casos, donde las características de los tumores son muy distintas y presentan distintas escalas. Además, al ser un modelo basado en el ensamblado de árboles de decisión aleatorios, es capaz de crear un modelo fuerte evitando el sobreajuste.

Para entrenar este modelo, al igual que en la Regresión Logística se han elegido a todas las variables, menos el ID, la Dimensión Fractal y la variable objetivo, como predictoras. Tras esto, se han divido los datos en entrenamiento y test con una proporción de 70%-30%, respectivamente. Tras esto se ha entrenado el modelo con 3 árboles, número que se ha concluido como óptimo tras varias pruebas.

C) Gradient Boosting

Este modelo se ha elegido debido a su capacidad para trabajar con datos desbalanceados y para capturar relaciones no lineales, que al igual que con Random Forest, es algo importante para los casos clínicos y los tratados en este proyecto en concreto. También se trata de un modelo que busca la optimización de forma aditiva, es decir que crea un modelo más fuerte y preciso mediante el aprendizaje secuencial de un conjunto de modelos más débiles, a diferencia del Random Forest, el cuál calcula el objetivo en base al promedio de los resultados de los árboles. Esto también ayuda a evitar el sobreajuste, aunque la dificultad está en elegir la combinación de parámetros óptima, debido a la flexibilidad del modelo.

Para construir el modelo, se ha realizado de manera similar a los anteriores, ya que se han dividido de la misma manera las variables predictoras y objetivo. También, al igual que en Random Forest, se ha divido los datos en 70%-30% para entrenamiento y test. Pero,

en cambio, a la hora de entrenar el modelo, las combinaciones de parámetros que se han encontrado óptimas han sido el uso de 40 modelos y de un learning rate de 0.01.

Interpretación y justificación de las métricas para la elección del modelo

Para comparar los tres modelos, se han evaluado mediante distintas métricas, las cuales se han elegido ya que son las más comunes y apropiadas para modelos de clasificación, y de esta manera se han observado las diferencias de rendimiento entre ellos eligiendo el mejor.

Tabla 3. Métricas de evaluación de los modelos de clasificación

Métrica	**Descripción**	**Fórmula**	**Regresión Logística**	**Random Forest**	**Gradient Boosting**
Accuracy	Porcentaje total de valores correctamente clasificados	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$	Train: 0.972	Train: 0.987	Train: 0.972
			Test: 0.976	Test: 0.935	Test: 0.935
Precision	Porcentaje de positivos que son realmente positivos	$\frac{TP}{(TP+TF)}$	Train: 0.972	Train: 0.987	Train: 0.973
			Test: 0.976	Test: 0.936	Test: 0.941
Recall	Positivos correctamente clasificados	$\frac{TP}{(TP+FN)}$	Train: 0.972	Train: 0.987	Train: 0.972
			Test: 0.976	Test: 0.935	Test: 0.935
F1-Score	Combinación de Precision y Recall	$\frac{(R*P)}{(R+P)}$	Train: 0.972	Train: 0.987	Train: 0.972
			Test: 0.976	Test: 0.935	Test: 0.934

Elaboración Propia

Como se observa en los valores de la tabla los tres modelos han presentado un rendimiento bastante alto en entrenamiento y test. Sin embargo, el que presenta mejores métricas es la regresión logística, ya que tiene los mayores valores en test, por lo que es el modelo que mejor generaliza. Esto puede deberse a que el random forest y el gradient boosting presenten un cierto grado de sobreajuste provocando una peor generalización a nuevos datos.

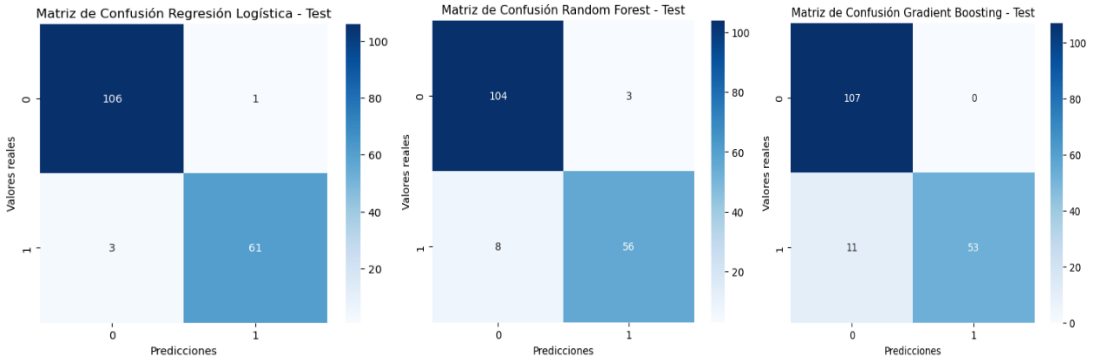

Ilustración 10. Matrices de confusión de los modelos de clasificación. Fuente: Elaboración propia con Python

Esto también se puede comprobar con las matrices de confusión donde se ve los 3 modelos presentan un gran rendimiento a la hora de diferenciar entre positivos y negativos reales. Sin embargo, Se observa que la mejor distinción ocurre en la regresión logística, donde únicamente hay 4 casos mal clasificados, mientras que en los otros dos modelos hay 11.

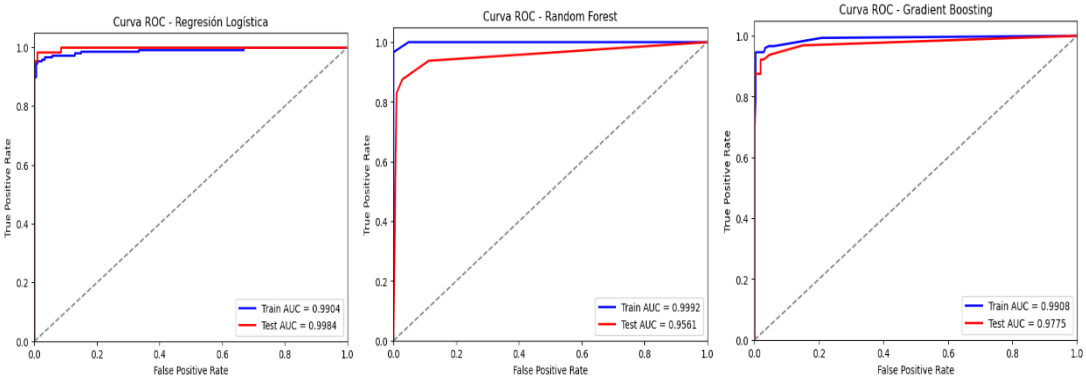

Ilustración 11. Curva ROC-AUC de los modelos de clasificación. Fuente: Elaboración propia con Python

Esto se corrobora con las curvas ROC-AUC presentando valores muy altos tanto en train como en test, pero siendo el de mayor train el Random Forest con el peor test, por lo

que se podría decir que generaliza peor, mientras que entre Gradient Boosting y Regresión Logística, el que presenta un mayor AUC para test, y que por tanto generaliza mejor, es la Regresión Logística.

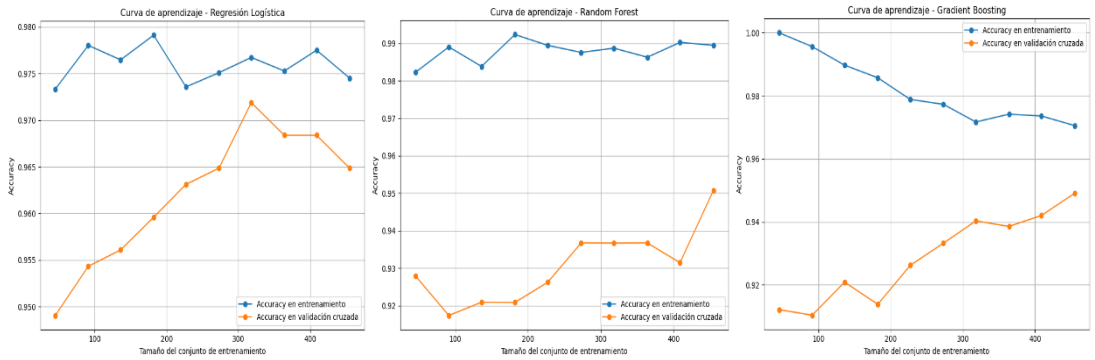

Ilustración 12. Curvas de aprendizaje de los modelos de clasificación. Fuente: Elaboración propia con Python

Finalmente, para comprobar el sobreajuste se han trazado las curvas de curvas de aprendizaje de cada modelo. Como se observa, el accuracy en train del Random Forest se mantiene prácticamente siempre en 0.99, mientras que en test prácticamente no aumenta hasta el final, por lo que está aprendido demasiado muy bien sobre los datos de entrenamiento, pero no está generalizando bien nuevos datos. En cambio, en el Gradient Boosting la curva en entrenamiento desciende progresivamente a la vez que la curva de validación aumenta, lo que quiere decir que el modelo aprende bien a la vez que adquiere una gran capacidad de generalización. Respecto a la Regresión Logística, se observa que la curva en entrenamiento presenta variaciones, por lo que va aprendiendo también progresivamente e incluso generaliza mejor que el Boosting, ya que la curva de validación consigue un pico mucho más alto.

Gracias a estas métricas y gráficos se ha podido comparar el rendimiento de los tres modelos concluyendo que el peor es el Random Forest, ya que es el que peor generaliza a nuevos datos, y que el mejor es la Regresión Logística, ya que es el modelo que mejores resultados presenta en la clasificación de nuevos datos. Sin embargo, el Gradient Boosting es un modelo con gran potencial, ya que no ha alcanzado su pico de precisión en cuanto a nuevos datos, pero sería necesario con más datos para ver si este se puede lograr.

Imágenes Ultrasonidos

Explicación del problema de análisis que se plantea

Se plantea un problema de clasificación de imágenes con el objetivo de predecir una clase en base al procesamiento de imágenes de ultrasonidos de tumores para poder realizar un diagnóstico médico. Se buscará identificar si una masa mamaria presenta un tumor benigno o maligno.

Para ello se ha planteado realizar un CNN mediante aprendizaje por transferencia. En concreto se utilizará la técnica de Partial Fine Tuning, ya que se pretende usar un modelo preentrenado descongelando únicamente las últimas capas. Tras el entrenamiento se hará una validación del modelo mediante distintas métricas para definir su capacidad para clasificar los tumores.

Justificación y detalle de los modelos

Se ha decidido realizar la CNN mediante aprendizaje por transferencia, ya que desarrollar uno desde cero requiere grandes volúmenes de datos y una gran capacidad computacional, con la que no se cuenta. Por ello se ha optado por esta técnica, ya que permite adaptar modelos ya preentrenados con mucho menor coste.

Para desarrollar esta técnica, se ha elegido el modelo DenseNet121, ya que presenta una arquitectura que destaca por su estructura de conexiones entre varias capas densas, permitiendo una mejor propagación del gradiente y reutilización de características. Además, se trata de una arquitectura que ha demostrado un gran rendimiento en tareas médicas y con datos limitados, como es este caso. Por otro lado, se ha elegido ya que, en comparación con otros modelos como ResNet50 o VGG19, ha presentado una mayor precisión a la hora de generalizar.

Como se ha comentado, tras probar con distintos modelos, como ResNet50, VGG19 y DenseNet121, y varias configuraciones con distintos learning rates, epochs, optimizadores y capas descongelas; el que mejor rendimiento ha mostrado ha sido DenseNet121. Para dicho modelo, primero se han establecido manualmente los pesos de las clases, para corregir el desbalanceo ya comentado, siendo 1 para benigno y 2 para maligno. En cuanto a la configuración del modelo, la que se ha encontrado óptima ha sido

un learning rate de 0.00001, con un optimizador Adam, una función de perdida Crossentropy, 200 epochs y 30 capas descongeladas. Se ha elegido este learning rate, ya que es el que mayor equilibrio ha presentado en el modelo evitando el overfitting y el underfitting. También se han elegido las funciones de optimización y perdida, ya que son las más comunes y adecuadas para problemas de clasificación. Respecto al número de epochs se ha elegido un número alto como es 200, ya que ofrece un amplio margen de aprendizaje durante el entrenamiento, a pesar de que se ha establecido un early stopping en el caso de que este no mejore. Y en cuanto al número de capas descongeladas, se han elegido 30, ya que tras varias pruebas se ha definido que es la cantidad necesaria para que el modelo se adapte a la tarea de clasificación de los tumores.

Además, para adaptar la estructura del modelo a la tarea especificada, se han añadido capas para la clasificación, en las que se han mantenido los pesos de ImageNet (base de datos de imágenes con la que ha sido preentrenado el modelo). Entre estas se ha añadido una capa de Global Average Pooling, seguida por una capa densa de 512 neuronas con una función de activación ReLu, una capa Dropout donde se ignorará al 50% de las neuronas para evitar el sobreajuste, y una capa de salida con una neurona y la función Sigmoidea para realizar la clasificación de los tumores en benignos y malignos. Además, como ya se han comentado se han descongelado las 30 últimas capas del modelo sin cambiar su estructura, pero con los pesos que se habían establecido manualmente.

También para mejorar la eficiencia del modelo se han incluido callbacks como ReduceLROnPlateau para que en el caso de que no mejore el modelo tras varios epochs se reduzca el learning rate a no más de 0.000001, EarlyStopping para detener el entrenamiento si el modelo tampoco mejora en varios epochs, y ModelCheckPoint para guardar los pesos únicamente cuando el modelo mejora.

Interpretación y justificación de las métricas

Tras entrenar el modelo, este se detuvo en el epoch 42, debido a que no mejoraba el accuracy de 82% en entrenamiento y 80% en validación. Además, el entrenamiento finalizó con un accuracy en test de 83%. Estas valores de accuracy son buenos para estos sets de datos y demuestran que el modelo generaliza bien al haber un valor alto para test sin tener una gran diferencia con train. En cuanto a la pérdida durante el entrenamiento se redujo desde un 1.05 hasta un 0.44 y en validación de un 0.73 a un 0.36, lo que quiere decir

que ha habido una buena progresión en el aprendizaje, ya que el terminó con una pérdida bastante baja en comparación con la del primer epoch. También finalizó con una pérdida de 0.42 en test, el cual también es in valor bastante bajo, incluso menor que el de train, lo que confirma que el modelo generaliza bien.

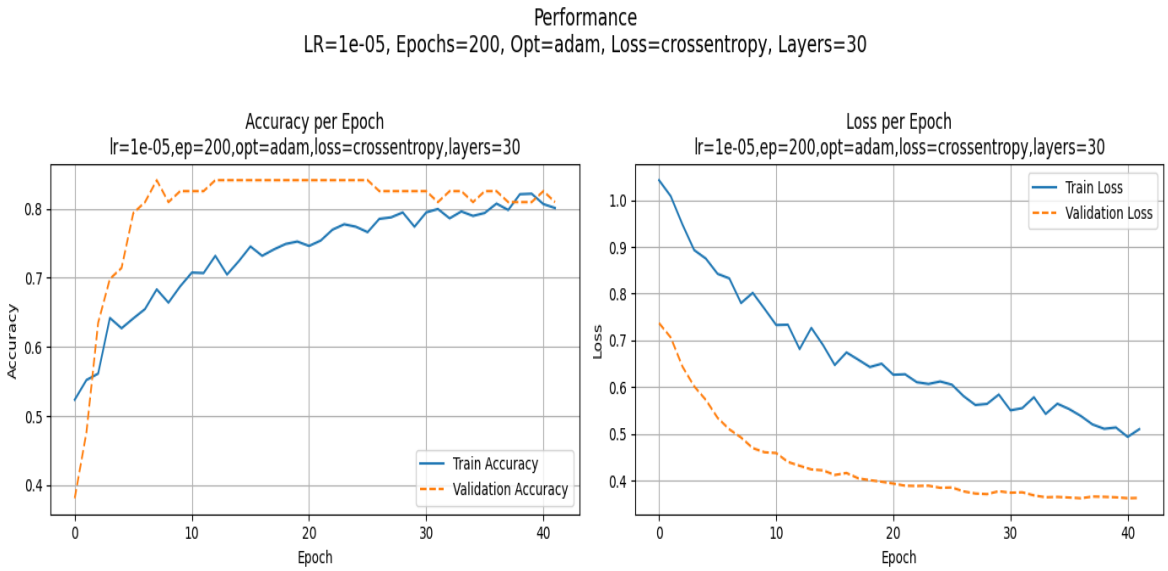

Ilustración 13. Curvas del Performance del modelo CNN. Fuente: Elaboración propia con Python

Lo comentado anteriormente se confirma observando estos gráficos. En el de la izquierda se encuentra un crecimiento del accuracy durante el entrenamiento bastante rápido en validación pero más progresivo en train, lo que indica que el modelo ha ido aprendiendo de forma correcta y sin caer en el sobreajuste. En el de la derecha, se observa algo similar al anterior, ya que en validación la pérdida disminuye más rápido que en train, que sigue siendo más progresivo. Esto confirma que el modelo no se ha sobreajustado durante el entrenamiento y generaliza bien sobre nuevos datos.

Sin embargo, el modelo se terminó de evaluar mediante las siguientes métricas en test para ambas clases:

Tabla 4. Métricas de evaluación de las clases del modelo CNN

Métrica	**Benigno**	**Maligno**
Precision	0.76	0.46
Recall	0.51	0.72
F1-Score	0.61	0.56

Elaboración Propia

Con estos valores se observa que el modelo presenta un rendimiento equilibrado, ya que detecta muy bien los tumores benignos, al tener una alta precisión, pero no detecta todos de forma correcta al tener un recall de 0.52. En cambio, los malignos se detectan mejor la mayoría al tener un recall alto, pero con una precisión baja, por lo que se pueden detectar falsos malignos. Esto puede deberse al thershold de 0.25, lo que quiere decir que si la probabilidad de que una pertenezca a una clase es mayor al 25% se le clasifica como maligna. Sin embargo, este threshold bajo fue elegido ya que es el que presenta mejor equilibro para clasificar imágenes nuevas. Además, para el ámbito médico es mejor detectar un tumor maligno que no lo sea a detectar como benigno a un tumor maligno, ya que de esta manera habría mucho menos peligro para los pacientes. De todas maneras, aunque generalmente el modelo clasifique bien ambas clases, se podría mejorar entrenándolo con más imágenes y así evitar una alta sensibilidad a una concreta siendo más imparcial y haciendo los diagnósticos más precisos.

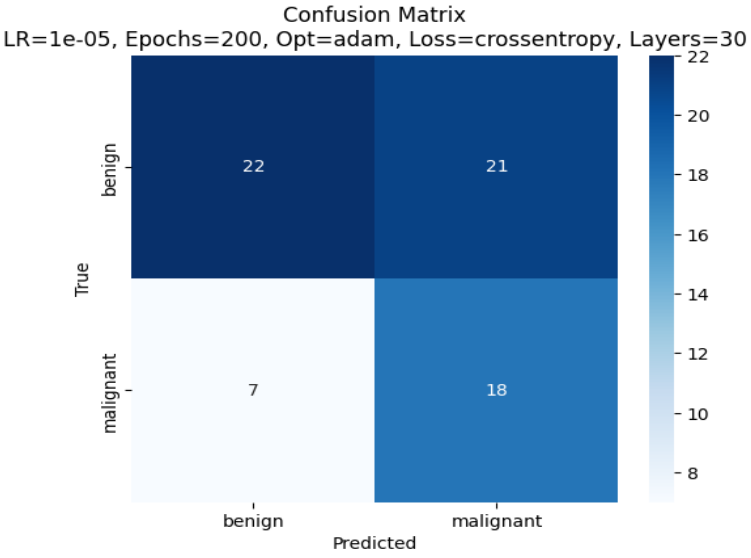

Ilustración 14. Matriz de confusión del modelo CNN. Fuente: Elaboración propia con Python

Esto también se puede ver en la matriz de confusión, ya que el modelo clasifica la mayoría de los tumores como malignos. Sin embargo, la mayoría de benignos y malignos verdaderos se clasifican de forma correcta. Esto indica la buena generalización del modelo, pero que también un potencial de mejora.

Visualización y discusión de los resultados obtenidos

Para comprobar los resultados, se han visualizado algunas de las imágenes del set de test y la probabilidad con la que el modelo las ha clasificado:

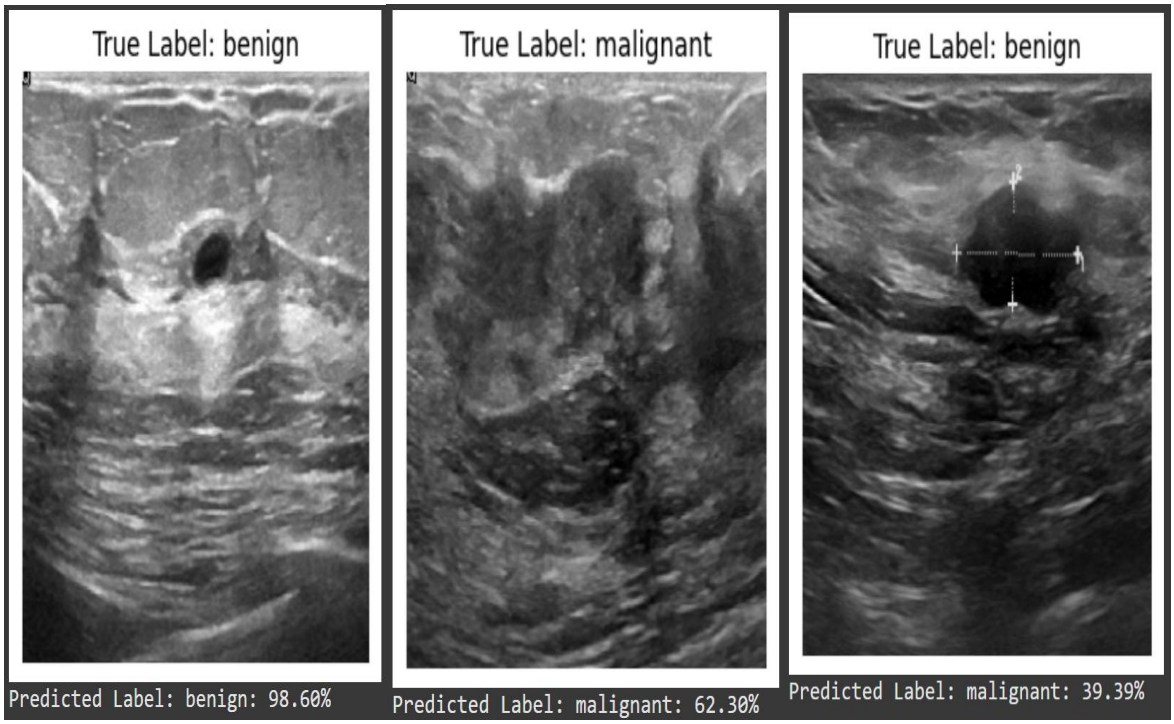

Ilustración 15. Ejemplos de imágenes clasificada por el modelo CNN. Fuente: Elaboración propia con Python

En estos ejemplos, se observa que a pesar del threshold, las imágenes clasificadas correctamente los hacen una probabilidad muy alta, incluso por encima del 50%. Sin embargo, las imágenes mal clasificadas, como la tercera, lo hacen con una probabilidad mucho más baja, aunque ligeramente superior al threshold de 25%. Esto quiere decir que el modelo aún es muy preciso, ya que lo que predice bien lo hace con mucha seguridad, mientras que lo que predice mal lo hace con probabilidades más bajas.

Esto es algo a tener en cuenta a la hora de aplicarlo ya que, si un médico observa que la clasificación tiene probabilidad alta, quiere decir que es correcta.

REFERENCIAS BIBLIOGRÁFICAS

Abbass, H. A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. Artificial Inteligence in Medicine.

Abeer Saber, A. G. (2023). Adapting the pre-trained convolutional neural networks to improve the anomaly detection and classification in mammographic images. Scientific Reports.

Amrane, M., Oukid, S., Gagaoua, I., & Ensarİ, T. (2018). Breast cancer classification using machine learrning. 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT).

Anthony Esteban Aldaz Noble, D. A. (2025). Análisis comparativo de redes neuronales para la predicción de cáncer de mama en imágenes médicas mediante la evaluación estadística de rendimiento. Obtenido de Repositorio Institucional de la Universidad Politécnica Salesiana : http://dspace.ups.edu.ec/handle/123456789/29913

Aprende Machine Learning. (2018). Breve Historia de las Redes Neuronales Artificiales. Obtenido de Aprende Machine Learning: https://www.aprendemachinelearning.com/breve-historia-de-las-redes-neuronales-artificiales/

Asociación Española Contra el Cáncer. (2023). Cáncer de mama. Obtenido de Contra el Cáncer: https://www.contraelcancer.es/es/todo-sobre-cancer/tipos-cancer/cancer-mama/prevencion/mamografias

Asociación española contra el cáncer. (2023). Causas y Factores de riesgo del cáncer de mama. Obtenido de Contra el cancer: https://www.contraelcancer.es/es/todo-sobre-cancer/tipos-cancer/cancer-mama/prevencion/factores-riesgo-cancer-mama

Asociación Española Contra el Cáncer. (2024). El cáncer de mama en España, en gráficos. Obtenido de epdata: https://www.epdata.es/datos/cancer-mama-espana-graficos/619/espana/106

Azar, A. T., & El-Said, S. A. (2013). Probabilistic neural network for breast cancer classification. Neural Computing and Applications.

Barajas, F. H. (2024). Modelos Predictivos. En F. H. Barajas, Gradient Boost.

Bergmann, D. (2024). ¿Qué es el fine-tuning? Obtenido de IBM: https://www.ibm.com/es-es/think/topics/fine-tuning

Chávez, C. E. (2023). Sistema automático para la interpretación inmediata de mamografías para la determinación del riesgo de cáncer. Obtenido de Universidad Autónoma de Aguascalientes: http://hdl.handle.net/11317/2583

Cox, D. (1958). The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society. Series B (Methodological).

DePolo, J. (2025). Tipos de cáncer de mama. Obtenido de BREASTCANCER.ORG: https://www.breastcancer.org/es/tipos

Friedman, J. H. (1999). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics.

Gabriel, R. P., López, V. R., & Barbosa, R. C. (2013). Redes Bayesianas para la clasificación de masas en mamografías. Universidad Tecnológica de la Mixteca.

Geeks for Geeks. (2024). Convolutional Neural Network (CNN) in Machine Learning. Obtenido de Geeks for Geeks: https://www.geeksforgeeks.org/convolutional-neural-network-cnn-in-machine-learning/

Geeks for Geeks. (2024). Geeks for Geeks. Obtenido de Logistic Regression in Machine Learning: https://www.geeksforgeeks.org/understanding-logistic-regression/

GeeksforGeeks. (2025). Bagging vs Boosting in Machine Learning. Obtenido de GeeksforGeeks: https://www.geeksforgeeks.org/machine-learning/bagging-vs-boosting-in-machine-learning/

GeeksforGeeks. (2025). Decision Tree in Machine Learning. Obtenido de GeeksforGeeks: https://www.geeksforgeeks.org/decision-tree-introduction-example/

GeeksforGeeks. (2025). Random Forest Algorithm in Machine Learning. Obtenido de GeeksforGeeks: https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., . . . Cardoso, F. (2019). Breast Cancer. Nature Reviews Disease Primers.

Harold Agudelo Gaviria, M. O. (2021). Detección de cáncer de seno usando imágenes de histopatología y modelos de aprendizaje profundo pre-entrenados. Journal of Computer and Electronic Sciences.

Hoss Belyadi, A. H. (2021). Gradient Boosting. Machine Learning Guide for Oil and Gas Using Python.

IBM. (s.f.). What is random forest? Obtenido de IBM: https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems

Khan, M. H.-M., Boodoo-Jahangeer, N., Dullull, W., Nathire, S., Gao, X., Sinha, G. R., & Nagwanshi, K. K. (2021). Multi- class classification of breast cancer abnormalities using Deep Convolutional Neural Network (CNN). Plos One.

Mayo Clinic. (2025). Cáncer de mama. Obtenido de Mayo Clinic : https://www.mayoclinic.org/es/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

MedilinePlus. (2024). Cáncer de mama. Obtenido de MedilinePlus: https://medlineplus.gov/spanish/ency/article/000913.htm

Melo, D. J. (2023). Redes Neuronales Convolucionales (Convolutional Neural Networks). Obtenido de Medium: https://medium.com/@djm9826/redes-neuronales-convolucionales-convolutional-neural-networks-6fb2f0bd9fe7

Pedro Moises de Sousa, L. A. (2024). Breast Density Classification Using Convolutional Neural Networks and Analysis of the CLAHE Technique. Obtenido de SBC OPEN LIB: https://doi.org/10.5753/wsis.2024.33673

Rivera, M. (2017). CNN con Aumentación de Datos. Obtenido de Centro de Investigación en Matemáticas CIMAT: http://personal.cimat.mx:8181/~mrivera/cursos/aprendizaje_profundo/aumentacion/aumentacion_datos.html

Tingting Liao, R. O. (2023). Classification of asymmetry in mammography via the DenseNet convolutional neural network. European Journal of Radiology Open.