

# Breast Cancer Detection

Joaquin Marroquin – Daniel Martinez  
20004254 - 19001064



## Introducción

El cáncer de mama es uno de los tipos de cáncer más comunes y una de las principales causas de muerte en mujeres a nivel mundial. La detección temprana y la clasificación precisa entre tumores malignos (cancerosos) y benignos (no cancerosos) son cruciales para el tratamiento adecuado y la mejora de la tasa de supervivencia.

En el siguiente proyecto, se desarrolló un modelo para clasificar tumores como malignos o benignos haciendo uso de datos con un diagnóstico, el uso de la inteligencia artificial tiene el potencial de reducir errores en el diagnóstico, disminuir la carga de radiólogos y patólogos, ofrecer varias opiniones en casos complejos, esto puede ayudar a detener el cáncer en etapas más tempranas, cuando es más tratable y las posibilidades de supervivencia son mayores.

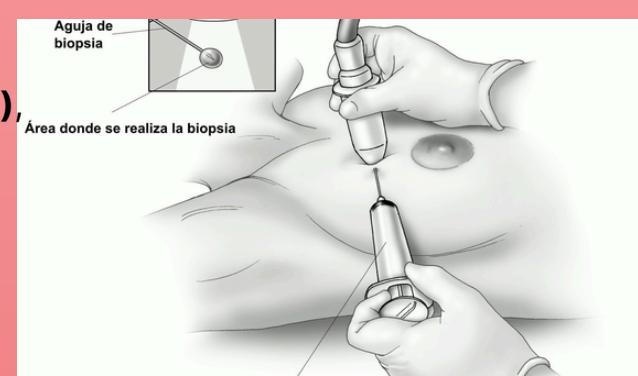
El conjunto de datos de cáncer de mama incluye información de 569 muestras de tumores, recolectadas mediante aspiraciones con aguja fina. Estas muestras se analizan y describen mediante 30 características, como el tamaño, la forma y la textura de las células.

- a) radio (media de las distancias desde el centro a los puntos del perímetro).
- b) textura (desviación estándar de los valores de la escala de grises).
- c) perímetro.
- d) área.
- e) suavidad (variación local en las longitudes de los radios).
- f) compacidad (perímetro / área).
- g) concavidad (severidad de las porciones cóncavas del contorno).
- h) puntos cóncavos (número de porciones cóncavas del contorno)
- y/o simetría.
- j) dimensión fractal (aproximación de la línea costera).

Para cada uno de estos parámetros se calcula la media (**mean**), error estándar (**se**) y peor caso (**worst**).

Los datos de los casos vienen identificados por dos tipos malignos y benignos:

Malignos: 212  
Benignos: 357



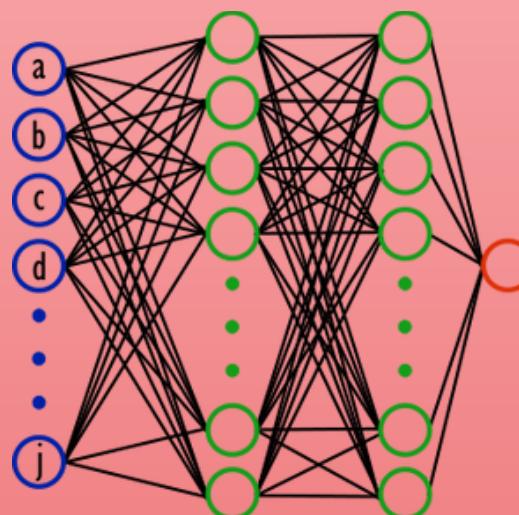
## Metodología

Para mejorar la calidad de los datos y facilitar el desarrollo de modelos de clasificación de cáncer de mama, se realizaron las siguientes modificaciones al conjunto de datos original:

- Se **normalizaron** los valores de las características utilizando **Z-score** para mantenerlos cercanos a cero y comparables entre sí.
- Codificar los valores de diagnóstico como 0 y 1 para representar maligno y benigno, respectivamente, en base a esto se realizó el primer método para obtener los resultados con la data desbalanceada.
- Dado el **desbalance** en la distribución de clases, con 357 muestras benignas y 212 malignas, se aplicó un segundo método de método de **sobremuestreo (oversampling)** para equilibrar el conjunto de datos, obteniendo 357 muestras de cada clase.
- Para el tercer método se aplicó **submuestreo (undersampling)** para equilibrar el conjunto de datos, obteniendo 212 muestras de cada clase.



## Modelo



El modelo está diseñado con:  
**input layer** – 30 entradas  
**hidden layers** – 2 capas de 16 neuronas cada una densamente conectadas  
**output layer** – 1 neurona

### Callbacks

**EarlyStopping** – tolerancia de 0.001 en accuracy para evitar sobreajustarse a los datos

Para la evaluación de nuestro modelo en la clasificación de tumores de mama como benignos o malignos, se realizaron 3 arquitecturas para tener una mejor visualización de los resultados, se mostró una alta precisión en la clasificación de los diagnósticos, esto indica que son efectivos tanto para identificar tumores malignos como para evitar falsos positivos en tumores benignos, la implementación de estos modelos puede mejorar la precisión del diagnóstico y asistir a los médicos en la toma de decisiones clínicas.

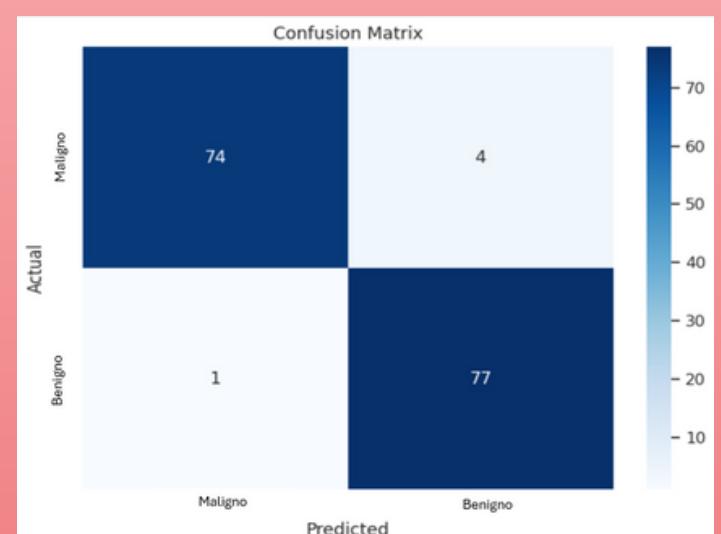
- **Modelo 1:** Se trabajó con el conjunto de datos desbalanceados, con valores entre diagnósticos entre 0 y 1 para que se tenga una mejor comparación entre los datos, al momento de la evaluación para este modelo se tuvo una exactitud del 95% teniendo 72 para el diagnóstico maligno y 76 para benigno, tomando los 155 datos que se tenían disponibles para hacer los test.
- **Modelo 2:** Para el trabajo con este modelo se realizó bajo la normalización entre diagnósticos principales, con los rangos de 0 a 1 de los parámetros importantes, para que el cesgo entre los datos no sea tan extenso, se balancearon los datos por medio de oversampling, aumentando genéricamente los datos de la clase más baja, en la evaluación se obtuvo una exactitud del 97% con 74 para el diagnóstico maligno y 77 para benigno, este resultado fue el más preciso.
- **Modelo 3:** Se realizó el balance de las clases por medio de undersampling que fue quitar datos de la clase de diagnóstico con más datos, obteniendo una exactitud del 89% siendo este nuestro modelo con menos precisión al clasificar datos, teniendo 67 maligno y 72 benigno

## Resultados

Modelo	Sensitivity	Specificity	Loss	Accuracy
Modelo 1	0.91	0.94	0.18	0.95
Modelo 2	0.93	0.96	0.16	0.97
Modelo 3	0.90	0.93	0.29	0.89

### Resultados de Evaluación:

Precisión General: 97%  
Diagnóstico Maligno: 74  
Diagnóstico Benigno: 77



## Mejoras a futuro

Un mejor desarrollo en la arquitectura del modelo para un mejor resultado, siendo así más comprensibles para el personal que necesita la asistencia.

Integración de más datos clínicos para una mejor precisión y robustez en el modelo.

Punto de aprobación clínico, validación y pruebas en entornos reales.

## Conclusiones

El uso de IA es efectivo en la clasificación de tumores, con un modelo de precisión del 97%, esto ofrece una herramienta poderosa para el área de salud en la detección temprana de un diagnóstico, lo que puede reducir los errores en uno, mejorando la atención y el pronóstico de los pacientes.