

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.

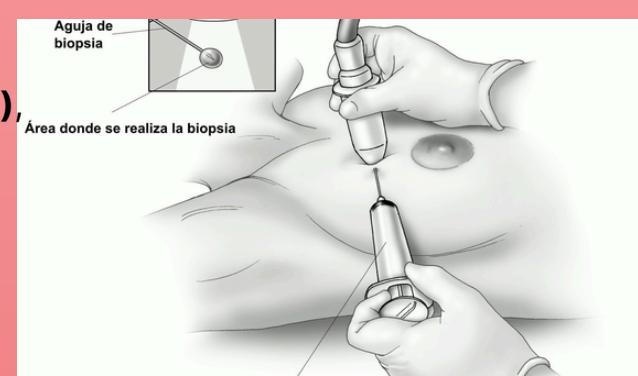
El conjunto de datos de cáncer de mama incluye información de 569 muestras de tumores, recolectadas mediante aspiraciones con aguja fina. Estas muestras se analizan y describen mediante 30 características, como el tamaño, la forma y la textura de las células.

- 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

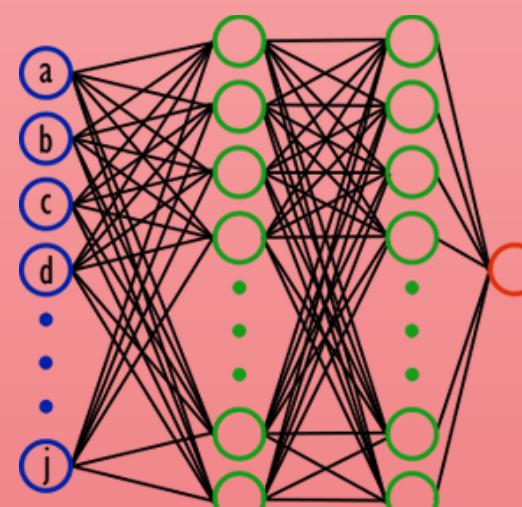
Para mejorar la calidad de los datos y facilitar el desarrollo de modelos de clasificación de cáncer de mama, se realizaron las siguientes modificaciones al conjunto de datos original:

- 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.

Para la evaluación de nuestro modelo en la clasificación de tumores de mama como benignos o malignos, se realizaron 3 procesamientos 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, a continuación se describen las diferencias entre los modelos.

- **Modelo 1:** Se trabajó con el conjunto de datos desbalanceados. Aplicando un preprocesamiento con las mismas modificaciones.
- **Modelo 2:** Para que el sesgo entre los datos no sea tan extenso, se **balancearon los datos por medio de OVERSAMPLING**, aumentando genéricamente los datos de la clase más baja, duplicando algunos de los datos ya existentes al azar.
- **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.

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



Resultados

Modelo	loss	accuracy	sensitivity	specificity
1	0.141266	0.957142	0.985714	0.928571
2	0.130415	0.964285	0.957143	0.971429
3	0.121814	0.949999	0.957143	0.942857

Resultados de Evaluación:
El mejor modelo fue el Modelo #2

Precisión General: 96.4%
Diagnóstico Maligno: 68
Diagnóstico Benigno: 67



Mejoras a futuro

- Un mejor desarrollo en la arquitectura del modelo para un mejor resultado.
- Integración de más datos clínicos para una mejor precisión y robustez en el modelo.
- Eliminación de datos irrelevantes o que no llevan relación con el diagnóstico.
- 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 96.4%, 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.
- Al momento de entrenar todos los modelos obtuvieron buenos resultados, a pesar de estar desbalanceados al tener suficientes datos no desvió la predicción.