# Análisis Exploratorio — Wisconsin Breast Cancer Diagnostic (WBCD)

Fase I (preprocesamiento) y Fase II (EDA)

## Descripción del dataset

El conjunto WBCD contiene 569 registros de mujeres y 32 columnas. La variable objetivo es 'diagnosis' (maligno=1, benigno=0). Cada observación incluye medidas geométricas y de textura de núcleos celulares.

## Fase I — Preprocesamiento

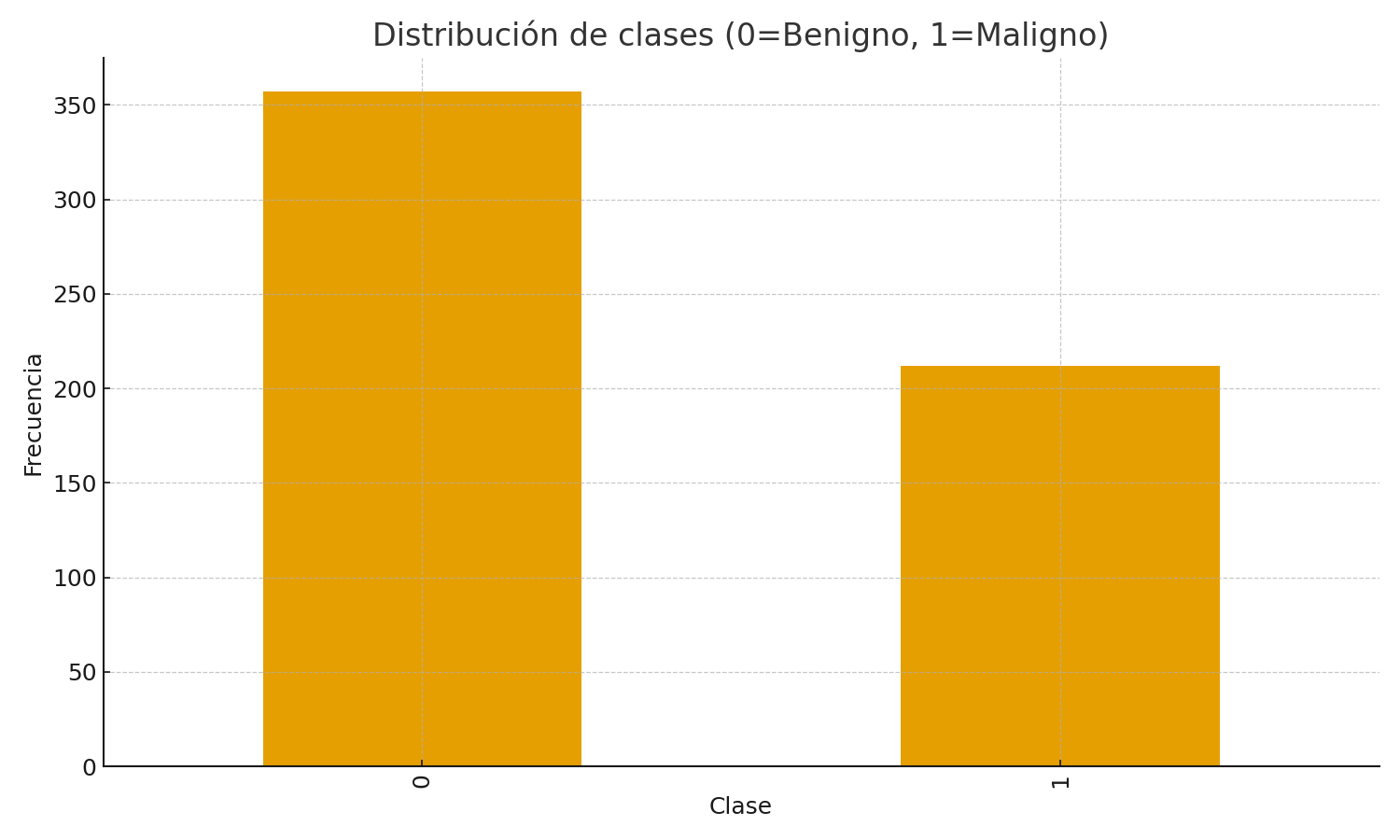
• Se sustituyeron etiquetas: M→1 y B→0 (maligno=1, benigno=0).

• Se descartó la primera columna (id) por no aportar al problema.

• Verificación de valores perdidos: 0 en total (no se detectaron valores perdidos).

• Distribución de clases: 0 (benigno) = 357, 1 (maligno) = 212.

Figura 1. Distribución de clases.



## Fase II — EDA (Resultados e interpretación)

Se calcularon estadísticas descriptivas y correlaciones con la variable objetivo. Las variables con mayor asociación absoluta con la etiqueta fueron:

1) worst concave points

2) worst perimeter

3) mean concave points

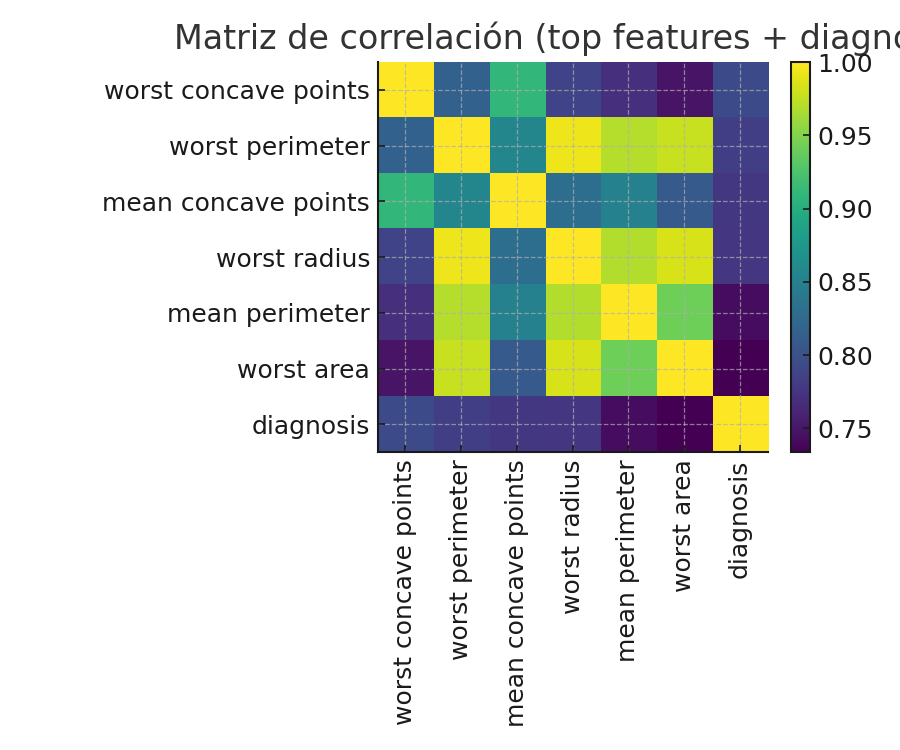
4) worst radius

5) mean perimeter

6) worst area

Los histogramas muestran colas más largas y valores mayores en pacientes malignos para medidas como 'worst radius', 'worst perimeter' y 'worst area'. Los boxplots por clase evidencian medianas más elevadas y mayor dispersión en la clase maligna, especialmente en 'concave points' (promedio y peor caso), indicando bordes más irregulares de los núcleos en tumores malignos.

Figura 2. Matriz de correlación (top features + diagnosis).



## Conclusiones

El preprocesamiento requerido se completó correctamente (codificación de la etiqueta, descarte de ID y verificación de nulos). El EDA sugiere que medidas relacionadas con tamaño, perímetro, área y, en particular, 'concave points', son altamente informativas para la predicción del diagnóstico. La distribución de clases está desbalanceada (357 benignos vs. 212 malignos), por lo que en una etapa de modelado convendría aplicar técnicas de estratificación y métricas robustas (ROC-AUC, sensibilidad) para evaluar correctamente el desempeño.

Archivos generados:

• WBCD\_clean.csv (dataset preprocesado).

• Figuras de EDA (histogramas, boxplots y correlación).