

# AUTO II PARTE 1- CAMILA ROA MONTOYA

## Identificación

### Instrucciones

#### 1. Análisis preliminar del problema

Para el dataset seleccionado:

- a. Determine si se trata de un problema de clasificación o regresión. Justifique su respuesta e indique claramente el target (variable objetivo).
  - b. Clasifique las características en tipos de variables (numéricas, categóricas, binarias, ordinales, etc.).
  - c. Investigue y explique el protocolo de adquisición y/o generación de datos que siguieron los investigadores.
- a. La variable objetivo o target del dataset es **diagnosis**, la cual indica si el tumor es **benigno (B)** o **maligno (M)**. Dado que esta variable es de tipo categórico con dos posibles resultados, el problema corresponde a una **clasificación**.
- b.

Tipo de Variable	Columnas
Numérica discreta	id (identificador)
Categórica binaria	diagnosis (M = maligno, B = benigno)
Numéricas continuas (float)	radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst,
Columna vacía	Unnamed: 32 (sin datos)

### c. Protocolo de adquisición de datos

Este conjunto de datos proviene del **Breast Cancer Wisconsin (Diagnostic) dataset**, recolectado en el *University of Wisconsin Hospitals, Madison* por el Dr. William H. Wolberg.

Los datos se obtuvieron a partir de **imágenes digitalizadas de muestras de tejido mamario**, tomadas mediante el procedimiento de **aspiración con aguja fina (FNA)**. A partir de estas imágenes, un sistema computarizado extrajo automáticamente **30 características numéricas** relacionadas con la forma, el tamaño y la textura de los núcleos celulares, como el radio, perímetro, área, concavidad o simetría.

El diagnóstico clínico de cada muestra fue determinado por especialistas, clasificándolas como **benignas (B)** o **malignas (M)**. El objetivo de este protocolo de recolección fue crear una base de datos que permitiera entrenar y evaluar modelos de aprendizaje automático capaces de apoyar el diagnóstico temprano del cáncer de mama.

#### Referencia:

Dua, D. & Graff, C. (2019). *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science. Disponible en: [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(diagnostic\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic))

investigación

#### K-Fold Cross Validation:

Es una técnica de validación en la que se divide el dataset en  $k$  partes iguales (folds). Se entrena el modelo  $k$  veces, cada vez usando  $k-1$  folds para entrenar y el fold restante para validar. Finalmente, se promedian las métricas obtenidas en cada iteración para tener una evaluación más robusta del desempeño del modelo.

#### Leave-One-Out Cross Validation (LOOCV):

Es un caso extremo de K-Fold donde  $k$  es igual al número de muestras del dataset. Es decir, se deja una muestra fuera para validación y se entrena el modelo con todas las demás, repitiendo el proceso para cada observación. Proporciona una estimación casi imparcial del error del modelo, pero es muy costoso computacionalmente para datasets grandes.

#### Aplicabilidad al dataset elegido:

Sí, ambas estrategias son aplicables al dataset. Dado que el dataset tiene 569 muestras, K-Fold (por ejemplo, con  $k=5$  o  $k=10$ ) es totalmente factible. LOOCV

también es posible, pero sería más lento porque requeriría entrenar el modelo 569 veces.

**Beneficios frente al esquema tradicional (train/val/test):**

- Proporcionan una estimación más **robusta y confiable** del desempeño del modelo, reduciendo la varianza debida a cómo se hace la división de los datos.
- Aprovechan **todas las muestras** tanto para entrenamiento como para validación en diferentes iteraciones, evitando que algunas observaciones “no contribuyan” a la evaluación.
- Permiten **comparar modelos** de manera más precisa, especialmente en datasets pequeños o medianos.