

## Introducción

El Breast Cancer Wisconsin Diagnostic Dataset es un conjunto de datos ampliamente utilizado en proyectos de ciencia de datos y aprendizaje automático, especialmente para prácticas de clasificación y agrupamiento. Fue recopilado por el Dr. William H. Wolberg en la Universidad de Wisconsin, y contiene información sobre características de células tumorales obtenidas a partir de imágenes digitales de biopsias de tejido mamario. El objetivo de estos datos es analizar las características físicas de los núcleos celulares para ayudar a distinguir entre tumores malignos y benignos.

## Objetivo General

Aplicar técnicas de **regresión, clasificación y redes neuronales** para explorar y analizar patrones en los datos del cáncer de mama, con el fin de:

1. Identificar relaciones entre variables mediante regresión lineal y logística.
2. Evaluar el desempeño de un modelo de red neuronal para clasificación.
3. Comparar resultados y probabilidades predichas por cada modelo.

## Objetivos Específicos

- Realizar **análisis exploratorio** del dataset.
- Implementar **regresión lineal multivariable** para predecir características continuas.
- Implementar **regresión logística** para clasificar tumores malignos y benignos.
- Diseñar y entrenar un **MLP en PyTorch** para clasificación binaria.
- Visualizar resultados y comparar desempeño entre métodos.

## Metodología

### 1. Exploración inicial del conjunto de datos

### 2. Regresión lineal multivariable

1. Selección de variables predictoras y variable objetivo continua (por ejemplo, `mean area`, `mean perimeter`).
2. Ajuste de modelo de regresión lineal multivariable.
3. Evaluación:
  - Coeficientes y su interpretación.
  - Error cuadrático medio (MSE).
4. Visualización:

- Predicciones vs valores reales.
- Tabla con algunas predicciones y sus errores.

### 3. Regresión logística

1. Selección de características relevantes para clasificación binaria.
2. Ajuste de un modelo de **regresión logística**:
  - Mostrar probabilidades de cada muestra de pertenecer a cada clase.
3. Evaluación:
  - Exactitud, matriz de confusión.
4. Visualización:
  - Probabilidades predichas.
  - Comparación de predicciones con clase real.

### 4. Red neuronal con PyTorch

1. Construcción de un **MLP** simple:
  - Input: número de características seleccionadas.
  - Capas ocultas con ReLU.
  - Output: 1 neurona con Sigmoid.
2. Entrenamiento:
  - Función de pérdida: BCE.
  - Optimizador: Adam.
  - Número de épocas y batch size.
3. Evaluación:
  - Exactitud en test.
  - Probabilidades predichas.
4. Visualización:
  - Curva de pérdida por época.
  - Comparación de predicciones con datos reales.
  - Subplots mostrando algunas predicciones correctas e incorrectas.

### 5. Comparación y análisis

1. Comparar **regresión logística y red neuronal**:
  - Exactitud.
  - Probabilidades predichas.
  - Ventajas y limitaciones de cada método.
2. Reflexión sobre cómo los modelos podrían aplicarse en un **contexto médico**.

## Preguntas reflexivas

1. ¿Qué modelo te pareció más adecuado para este conjunto de datos y por qué?
2. ¿Qué criterios utilizaste para seleccionar las variables predictoras?
3. ¿Cómo influyó el número de capas y neuronas en la red neuronal en el desempeño del modelo?
4. ¿Qué ventajas observaste al usar redes neuronales frente a regresión logística?
5. ¿Cómo podrían usarse los resultados de los modelos en decisiones médicas o investigación clínica?

## Entregables

- Notebook de Jupyter (.ipynb) con:
  - Código comentado de regresión lineal, regresión logística y red neuronal.
  - Visualizaciones y tablas de predicciones.
- Informe PDF:
  - Explicación del análisis realizado.
  - Resultados comparativos.
  - Respuestas a las preguntas reflexivas.

### Rúbrica de evaluación

Criterio	Descripción	Puntaje Máximo
1. Exploración inicial de datos	Carga correcta del dataset, revisión de estructura, tipos de datos, valores faltantes, y estadísticas básicas.	10
2. Selección y justificación de variables	Elección de variables adecuada y justificación razonada. Inclusión de visualización inicial.	10
3. Regresión lineal multivariable	Implementación correcta y análisis de coeficientes y error.	10
4. Regresión logística	Implementación correcta, exactitud y probabilidades de cada clase.	10
5. Red neuronal PyTorch	Implementación correcta, entrenamiento, curvas de pérdida y exactitud.	15
6. Comparación de resultados	Comparación entre regresión logística y red neuronal, con análisis de ventajas y limitaciones.	15
7. Visualización de resultados	Subplots de predicciones correctas e incorrectas, y gráficas claras.	10
8. Respuestas reflexivas	Respuestas completas y coherentes a las preguntas asignadas.	10
9. Presentación del informe o notebook	Claridad en formato, títulos, etiquetas y comentarios.	5
10. Ortografía, redacción y estilo	Buena redacción y uso técnico adecuado.	5