Ejercicio de repaso: modelado supervisado en problemas de salud de clasificación binaria

Contenido

[Recursos útiles 1](#_Toc207036494)

[Enunciado 2](#_Toc207036495)

[Parte I – Breast Cancer 2](#_Toc207036496)

[Paso 1: Carga y exploración inicial 2](#_Toc207036497)

[Paso 2: EDA básico 2](#_Toc207036498)

[Paso 3: Preparación del modelo 2](#_Toc207036499)

[Paso 4: Evaluación del modelo 2](#_Toc207036500)

[Parte II – Heart Disease (UCI) 3](#_Toc207036501)

[Paso 1: Carga y exploración inicial 3](#_Toc207036502)

[Paso 2: EDA básico 3](#_Toc207036503)

[Paso 3: Preprocesamiento 3](#_Toc207036504)

[Paso 4: Modelado 3](#_Toc207036505)

[Paso 5: Evaluación final 3](#_Toc207036506)

# Recursos útiles

Recuerde el directorio de recursos compartidos disponible en:

<https://nascorformacion0-my.sharepoint.com/:f:/g/personal/juan_pinuela_docente_nascorformacion_com/El63bU3PsG5CuybEDz8vH04BL1J5O88yNhhYMNYynPpZWA?e=k5slLd>

# Enunciado

Realizar un flujo completo de trabajo en *Machine Learning supervisado*, desde un análisis exploratorio de datos (EDA) hasta la construcción y evaluación de modelos predictivos de clasificación. Se abordarán dos datasets de referencia en el ámbito de la salud:

# Parte I – Breast Cancer

## Paso 1: Carga y exploración inicial

* Importar el dataset `load\_breast\_cancer()` desde `sklearn.datasets`.
* Convertirlo en un `DataFrame` con nombres de columnas.
* Analizar distribución de la variable objetivo (proporción benigno vs maligno).

## Paso 2: EDA básico

* Describir estadísticamente las variables (media, desviación típica, min/max).
* Visualizar correlaciones entre variables numéricas (heatmap).
* Comparar gráficamente algunas variables relevantes según la clase objetivo (ej. `mean radius` o `mean texture`).

## Paso 3: Preparación del modelo

* Realizar un train/test split estratificado.
* Construir un pipeline con:
  + `StandardScaler` para estandarizar.
  + Clasificador (`LogisticRegression` o `RandomForestClassifier`).

## Paso 4: Evaluación del modelo

* Validación cruzada con métricas: accuracy, precision, recall y f1.
* Entrenar en train, evaluar en test:
  + Matriz de confusión.
* Interpretación de resultados:
  + Importancia de coeficientes (LR).
  + Importancias de características (RF).

# Parte II – Heart Disease (UCI)

## Paso 1: Carga y exploración inicial

* Descargar dataset `fetch\_openml('heart-disease-uci')`.
* Normalizar la variable objetivo a binaria (0 = no enfermedad, 1 = enfermedad).
* Identificar columnas categóricas (ej: `sex`, `cp`, `thal` …).

## Paso 2: EDA básico

* Distribución de la variable objetivo (equilibrio de clases).
* Tablas cruzadas de variables categóricas vs target.
* Boxplots de variables numéricas según presencia/ausencia de enfermedad.
* Detección de valores perdidos o codificados como “?”.

## Paso 3: Preprocesamiento

* Usar un ColumnTransformer con dos ramas:
  + Numéricas → imputación (mediana) + escalado (`StandardScaler`).
  + Categóricas → imputación (moda) + `OneHotEncoder`.

## Paso 4: Modelado

* Definir pipelines con el preprocesador y un clasificador (`LogisticRegression`, `RandomForestClassifier`).
* Realizar validación cruzada con las mismas métricas.

## Paso 5: Evaluación final

* Entrenar en train, evaluar en test con métricas, matriz de confusión y curvas ROC/PR.
* Interpretar coeficientes de LR e importancias de RF.