

**EXAMEN FINAL DEL 3 DE FEBRERO DE 2024**  
**CASO PRACTICO MEDIANTE EL USO DE PYTHON**  
**Tiempo: 1 hora**

**Apellidos:**

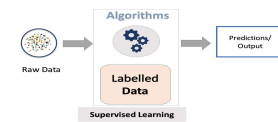
**Nombre:**

**Firma:**

La Cleveland Heart Disease Database (CHDD) se considera la base de datos de referencia para la investigación de enfermedades cardíacas. El fichero dataset.csv contiene la información. Las variables incluidas son las siguientes.

- 1) Age. En este conjunto de datos se han seleccionado pacientes de edades comprendidas entre 29 y 79 años.
- 2) Sex. Los pacientes varones se denotan con un valor de género 1 y las pacientes mujeres con un valor de género 0.
- 3) Cp. Cuatro tipos de dolor torácico pueden considerarse indicativos de enfermedad cardíaca. La angina de tipo 1 está causada por una reducción del flujo sanguíneo a los músculos del corazón debido al estrechamiento de las arterias coronarias. La angina de tipo 1 es un dolor torácico que se produce durante situaciones de estrés mental o emocional. El dolor torácico no anginoso puede deberse a diversos motivos y no suele deberse a una cardiopatía real. El cuarto tipo, Asintomático, puede no ser un síntoma de enfermedad cardíaca.
- 4) Trestbps. Es la lectura de la tensión arterial en reposo.
- 5) Chol. Es el nivel de colesterol.
- 6) Fbs. Es el nivel de azúcar en sangre en ayunas; el valor se asigna como 1 si el azúcar en sangre en ayunas es inferior a 120 mg/dl y 0 si es superior.
- 7) Restecg. Es el resultado electrocardiográfico en reposo.
- 8) Thalach. Es la frecuencia cardíaca máxima.
- 9) Exang. Es la angina inducida por el ejercicio que se registra como 1 si hay dolor y 0 si no hay dolor.
- 10) Oldpeak. Es la depresión del ST inducida por el ejercicio.
- 11) Slope. Es la pendiente del pico del segmento ST del ejercicio.
- 12) Ca. Es el número de vasos principales coloreados por fluoroscopia.
- 13) Thal. Es la duración de la prueba de esfuerzo en minutos.
- 14) Target. Es el atributo de clase. El atributo de clase tiene un valor de 0 para normal y 1 para pacientes diagnosticados de cardiopatía.

Mediante el desarrollo en código PYTHON, proceder a generar un fichero que dé respuesta a cada apartado.



a) Construye el diagrama de dispersión y la recta de regresión lineal de Thalach sobre Age. Explica si es predecible la frecuencia cardíaca a partir de la edad, indicando el modelo con la estimación de sus parámetros, sus errores y su relevancia, incluyendo la tabla correspondiente. (2,5 puntos). Ayuda.- Prácticas 1 y 2.

Ante la problemática de decidir si un paciente tiene cardiopatía o no, se toma como conjunto de entrenamiento (TR) un 69 % de filas al azar (son 209 filas) y un conjunto de test de un 31% restantes de filas (son 94 filas).

b) Entrenar un modelo de regresión logística para predecir la probabilidad de que un paciente tenga cardiopatía basándose en la variable Restecg. Idem basándose en la variable Age. Adjuntar los gráficos correspondientes con su interpretación. Comparar la matriz de confusión en TR con la matriz de confusión en TS. (3,25 puntos) Ayuda.- Práctica 4.

c) Entrenar de manera razonada un árbol de clasificación o decisión para decidir si un paciente tiene cardiopatía basándose en las variables más relevantes. Adjuntar los gráficos correspondientes con su interpretación. Comparar la matriz de confusión en TR con la matriz de confusión en TS (3,25 puntos) Ayuda.- Práctica 5.

d) ¿Se trata de un conjunto de datos desbalanceados respecto a la presencia o no de cardiopatía? Razonar si en este caso es preferible la regresión logística o el árbol de clasificación para decidir si un paciente tiene cardiopatía. (1 punto).