

# Práctica 2

# Enfermedad

# Cardiovascular

Análisis de Datos

Grado en Ingeniería Informática

Curso 2020/21





# Enfermedad cardiovascular

- Queremos construir un sistema de apoyo al diagnóstico, que trate de «predecir» si un paciente tiene o no una enfermedad cardiovascular
- 70.000 pacientes
- 11 variables médicas, de tres tipos:
  - Objetivas
  - Resultantes de un examen médico
  - Subjetivas (indicadas por el paciente)
- Salida binaria (**problema de clasificación**)

# Variables

age	edad del paciente en días	objetiva
height	altura del paciente en cm.	objetiva
weight	peso del paciente en kg.	objetiva
gender	sexo biológico del paciente	objetiva
ap_hi	presión sistólica en mmHg	resultado de examen
ap_lo	presión diastólica en mmHg	resultado de examen
cholesterol	nivel de colesterol en sangre	resultado de examen
gluc	nivel de glucosa en sangre	resultado de examen
smoke	si el paciente es fumador habitual	subjetiva
alco	si el paciente es consumidor habitual de alcohol	subjetiva
active	si el paciente realiza ejercicio con regularidad	subjetiva
cardio	presencia o ausencia de enfermedad cardiovascular	

# Objetivo

Queremos construir un modelo de clasificación que pueda servir de herramienta de apoyo al diagnóstico, con el fin de ayudar a un especialista a tratar de pronosticar el riesgo de enfermedad cardiovascular dada cierta información básica de un paciente.

Además, el modelo entrenado debe poder interpretarse con el fin de determinar qué variables son más relevantes en el diagnóstico.

A continuación se describen las tareas para alcanzar este objetivo de forma satisfactoria.

# Tarea 1

## Entrenamiento y evaluación de un árbol de decisión | 4 puntos

En Aula Global se proporciona un fichero CSV con los datos.

Se deben emplear estos datos para construir un árbol de decisión, haciendo uso de todas las variables disponibles.

Una vez entrenado el árbol de decisión, deben proporcionarse métricas objetivas de su rendimiento, y deben discutirse. Los estudiantes deben decidir las métricas que estiman más oportunas para comunicar los resultados.

**Nota:** para este entrenamiento y evaluación se emplearán todos los datos disponibles, por lo que la evaluación se llevará a cabo sobre el propio conjunto de entrenamiento.

# Tarea 2

## Inspección del árbol de decisión | 2 puntos

La razón de escoger un árbol de decisión en la tarea anterior es sencilla: es un modelo que se presta bien a una inspección por parte de un especialista.

Empleando únicamente el modelo entrenado en la tarea anterior, tratar de dar respuesta a las siguientes preguntas:

- I. ¿Qué relevancia tienen los exámenes clínicos realizados a la hora de determinar la existencia de una enfermedad cardiovascular?
2. ¿Es relevante la información (subjetiva) proporcionada por los pacientes a la hora de determinar la presencia de una enfermedad cardiovascular?

# Tarea 3

## Comparativa de modelos | 4 puntos

Se deben entrenar otros modelos de clasificación (incluso si no son interpretables al nivel de un árbol de decisión) y compararse. Este proceso se realizará empleando validación cruzada con siete hojas (10.000 pacientes cada hoja).

Además, siempre que sea posible, queremos probar varios hiperparámetros (atributos de configuración) distintos para cada una de las técnicas.

Si se desea, se pueden probar preprocesamientos adicionales (*one-hot encoding*, normalización de los datos, etc.), siempre que estén justificados (es decir, si se hace, debe explicarse por qué se considera que es un proceso razonable).

El proceso de comparación de soluciones puede automatizarse si se considera oportuno.

# Tarea Extra

Estudio de los datos | +1 punto

En esta tarea **opcional** se permitirá a los estudiantes completar las respuestas a la Tarea 2 realizando un análisis más exhaustivo de los datos.

Es decir, se podrán realizar análisis específicos (por ejemplo, de correlación, de eliminación de variables, etc.) con el fin de dar respuesta a estas preguntas, sin necesidad de inspeccionar el árbol de decisión de forma directa.

# Evaluación

Cada tarea será evaluada con una puntuación, cuyo valor máximo viene indicado en las diapositivas anteriores.

Sean  $T_1, \dots, T_3$  las puntuaciones obtenidas por un equipo de prácticas en las tres primeras tareas, y  $T_E$  la puntuación obtenida en la tarea extra opcional. La nota de la práctica se calculará como:

$$P = \min(T_1 + T_2 + T_3 + T_E, 10)$$

Por lo tanto, la máxima nota alcanzable en la práctica es un 10.

Para evaluar cada tarea, se tendrá en cuenta que se haya implementado el código requerido, y que se hayan explicado las decisiones tomadas y los resultados obtenidos.

No es fundamental saber explicar todos los resultados de forma precisa, pero sí se debe intentar entenderlos.



# Entrega

- Se debe trabajar **individualmente o en parejas**.
- El trabajo vale 1 punto sobre la evaluación continua.
- **¡Importante!** Se debe revisar que todo el código funcione.
  - Es útil, antes de entregar la práctica, reiniciar el entorno y ejecutar todas las celdas.
- **Entrega:**
  - Viernes 13 de noviembre a las 23:59 a través de Aula Global.
  - Se entregará el cuaderno en formato IPYNB (y, si es posible, también en PDF).