Departamento de Ciencia de la Computación e Inteligencia Artificial
Universidad de Alicante **Curso 2025/2026**

Práctica 2. Árboles de decisión

Descripción general

En esta práctica vamos a analizar el funcionamiento de árboles de decisión para realizar clasificación de patrones. Partiremos de un conjunto extenso de datos con el que podremos reconocer si una persona tiene diabetes y de qué tipo es. Por lo tanto, la práctica se convierte en un proceso de aprendizaje automático supervisado en el que utilizaremos como modelo de aprendizaje un árbol de decisión.
Para el desarrollo de la práctica se utilizará Python como lenguaje de programación.

Base de datos

Para la realización de la práctica se proporciona el fichero diabetes_dataset.csv que contiene la información con la que vamos a entrenar el árbol de decisión. El fichero incluye 100.000 ejemplos etiquetados.
Los atributos son

- Age: Numérico 18-90
- Gender: Male, Female, Other
- Ethnicity: Asian, White, Hispanic, Black, Other
- Education level: No Formal, Highschool, Graduate, Postgraduate
- Income Level: Low, Lower, Lower Middle, Middle, Upper Middle, High
- Employment Status: Employed, Unemployed, Retired
- Smocking Status: Never, Former, Current
- Alcochol consumption per week: 0, 1, 2, 3, 4, 5, 6, 7
- Physical activity minutes per week: Numérico
- Diet Score: Numérico (0-10)
- Sleep hours per day: Numérico
- Screen time hours per day: Numérico
- Family shitory diabetes: 1, 0
- Hypertension History: 1, 0
- Cardiovascular history: 1, 0
- MBI: Numérico
- Waist to hip Ratio: Numérico (0-1)
- Systolic BP: Numérico
- Diastolic BP: Numérico
- Heart Rate: Numérico
- Cholesterol Total: Numérico
- HDL Cholesterol: Numérico
- LDL Cholesterol: Numérico
- Triglycerides: Numérico
- Glucose Fasting: Numérico
- Glucose Postprnadial: Numérico

- Insulin Level: Numérico
- HBA1C: Numérico

La **clase** que vamos a intentar aprender con el árbol de decisión es *Diabetes Stage* que puede tomar los valores: No Diabetes, Pre Diabetes, Type 1, Type 2.
El fichero contiene dos columnas: diabetes_risk_score y diagnosed_diabetes que no vamos a utilizar.

Actividad 1. Organizar las categorías de los atributos (2 puntos)

Como hemos visto, el fichero que contiene los datos de entrada tiene multitud de atributos. Algunos de ellos son categóricos y otros son numéricos. Estos últimos no se pueden utilizar tal cual están con árboles de decisión. Por lo tanto, la primera actividad consistirá en crear un primer abrá que justificar en la documentación el resultado de la categorización utilizando la teoría de la información.

Actividad 2. Entrenar un árbol de decisión (5 puntos)

El objetivo de esta actividad es crear un árbol de decisión a partir del fichero de datos de entrada mediante el algoritmo ID3 explicado en clase. Tendremos que encontrar el mejor árbol posible, es decir, el que clasifique mejor. Para ello haremos uso de un conjunto de datos para entrenamiento y otro para validación. Seguiremos el proceso que se ha explicado en clase. Deberemos documentar las estructuras que hemos creado para representar el árbol y el resultado de la clasificación con el árbol que hemos obtenido

Actividad 3. Entrenar con pre-poda (3 puntos)

En esta actividad buscamos mejorar los resultados de clasificación del árbol de la actividad anteriore. Para ello crearemos una variante del algoritmo ID3 en el que utilicemos alguno de los métodos de pre-poda vistos en clase. Documentaremos el método de pre-poda implementado y compararemos los resultados con los de la actividad anterior.

Entrega de la práctica

La fecha límite de entrega de los dos ejercicios de la primera práctica será el día 10 de enero de 2026. La entrega se realizará a través de una tarea en la página moodle de la asignatura. Los alumnos deberán entregar un fichero comprimido que contenga los ficheros fuente que hayan creado y el fichero con la documentación en **formato PDF**.

- ¡¡¡IMPORTANTE!!! Recordad que las prácticas son individuales y NO se pueden hacer en parejas o grupos. Cualquier código copiado supondrá un suspenso de la práctica para todas las personas implicadas en la copia. Se utilizarán herramientas para la detección de copia. Estos hechos serán comunicados a la Escuela Politécnica para que se tomen las medidas disciplinarias oportunas contra los infractores.