UNIVERSITAS
Miguel Hernández

Tema 5. Arboles de Decisión. Práctica.

José L. Sainz-Pardo Auñón

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE II

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones.

Descarga y Carga de los Datos

- Descarga el dataset **Breast Cancer Wisconsin (Diagnostic)** desde el UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`.
- Carga los datos en tu entorno de programación utilizando Python. Asegúrate de que todas las variables se carguen correctamente.
- Revisa las primeras filas del dataset y obtén un resumen estadístico básico para entender la distribución de las características.

Explora los Datos

- Analiza la variable objetivo (`Diagnosis`) y asegúrate de identificar la proporción de casos benignos y malignos.
- Visualiza las distribuciones de las variables más importantes como `Radius Mean`, `Texture Mean`, y otras características clave.
- Convierte la variable `Diagnosis` a numérica: asigna 0 a los valores "B" (benigno) y 1 a los valores "M" (maligno).
- Dibuja un mapa de correlaciones entre las variables para identificar relaciones importantes.

Preprocesa los Datos

- Elimina la columna `ID`, ya que no aporta información relevante al modelo.
- Divide el dataset en conjunto de entrenamiento (70%) y de prueba (30%) utilizando la función `train_test_split`.

Ajusta el Modelo de Arboles de Decisión

- Obtén varios árboles de decisión configurando distintos parámetros (prueba manualmente y con un grid de parámetros de búsqueda).
- Visualiza sus esquemas de decisión. Interprétalos. ¿Qué árbol es el que consideras más relevante?
- Obtén los pronósticos de la muestra de prueba con los parámetros del árbol que consideres más relevante.
- Obtén la tabla de confusión y el informe de clasificación del modelo, utilizando la librería sklearn.

Practica con Otros Datasets

- Repite la práctica con otros datasets disponibles en el UCI Machine Learning Repository y ya conocidos tanto de Regresión como de Clasificación. Por ejemplo:
  - **Pima Indians Diabetes Database:** `https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes`
  - **Heart Disease Dataset:** `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`
  - **Bank Marketing Dataset:** `https://archive.ics.uci.edu/ml/datasets/Bank+Marketing`
  - **Adult (Census Income) Dataset:** `https://archive.ics.uci.edu/ml/datasets/Adult`
  - **Auto MPG Dataset:** `https://archive.ics.uci.edu/dataset/9/auto+mpg`
  - **Wine Quality Dataset:** `https://archive.ics.uci.edu/dataset/186/wine+quality`
  - **Energy Efficiency Dataset:** `https://archive.ics.uci.edu/dataset/242/energy+efficiency`
  - **Concrete Compressive Strength:** `https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength`
- Sigue el mismo flujo de trabajo: carga los datos, explora las variables, preprocesa el dataset, ajusta el modelo y evalúa los resultados sin y con validación cruzada. Prueba también a seleccionar prototipos.
- Compara los resultados con aquellos obtenidos con Regresión Lineal, Logística y/o kNN.