TRABAJO PRÁCTICO 04: Árboles de decisión

Introducción

En este trabajo práctico veremos cómo ajustar árboles de decisión y cómo funcionan los principales algoritmos a través de evaluar diferentes parámetros.. En primer lugar, se utilizarán las nociones de entropía y ganancia de información introducidas en clase a efectos de entender cómo son los mecanismos para generar un árbol de decisión a partir de un dataset. Luego, se utilizará el lenguaje Python, puntualmente la librería Scikit-Learn.

Consignas

1. Con el siguiente dataset, implementar en python dos funciones, una que realice el cálculo de entropía y otra que calcule la ganancia de información. A partir de esas funciones genere el árbol de decisión para el dataset que se muestra a continuación:

```
PRONÓSTICO,TEMPERATURA,HUMEDAD,VIENTO,ASADO
Soleado,Calor,Alta,leve,no
Soleado,Calor,Alta,fuerte,no
Nublado,Calor,Alta,leve,si
Lluvioso,templado,Alta,leve,si
Lluvioso,Frío,normal,leve,si
Lluvioso,Frío,normal,fuerte,no
Nublado,Frío,normal,fuerte,si
Soleado,templado,Alta,leve,no
Soleado,Frío,normal,leve,si
Lluvioso,templado,normal,leve,si
Soleado,templado,normal,fuerte,si
Nublado,templado,Alta,fuerte,si
Nublado,Calor,normal,leve,si
Lluvioso,templado,Alta,fuerte,no
```

2. Trabaje con el dataset Breast Cancer Wisconsin (Diagnostic)[1].
    a. Utilice el metadata que provee la librería, ¿Cuál es el tema que aborda el dataset?
    b. Genere el árbol de decisión que permita clasificar los diagnósticos utilizando un muestreo con proporciones de 80% para entrenamiento y 20% para testeo.
    c. Explore la solución dada y las posibles configuraciones para obtener un nuevo árbol que clasifique "mejor". Documente las conclusiones.

3. Trabaje con el dataset de Pima Indians Diabetes[2].
    a. Utilice los metadatos disponibles en Kaggle para entender los datos, ¿Cuál es el tema que aborda el dataset?
    b. ¿Qué variable se debe modelar? ¿Cómo se distribuye? ¿Está balanceada?
    c. Genere el árbol de decisión que permita modelar la variable objetivo utilizando como criterio la entropía. Defina una estrategia de hold-out con proporciones de 80% para entrenamiento y 20% para testeo. Evalúe el resultado obtenido.
    d. Ahora, evalúe diferentes combinaciones de valores para los parámetros:
        i. max_depth
        ii. min_samples_split
        iii. min_samples_leaf
        iv. max_leaf_nodes

        Realice al menos 10 corridas y tabule las configuraciones con los respectivos resultados.
    e. Repita el experimento en d) agregando una validación cruzada con 5-folds
    f. Documente las conclusiones.

[1] https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
[2] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database