Boletín 3: árboles

Para la realización de las prácticas correspondientes a este boletín se utilizará scikit-learn.

1. Dado el siguiente conjunto de datos de clasificación con 6 observaciones, 3 variables de entrada y una variable de salida:

Observación	X_1	X_2	X_3	Y
1	4	3	-1	1
2	-3	-1	-1	0
3	3	-2	0	0
4	1	4	0	1
5	-2	3	1	0
6	-3	5	5	0

   Construye el árbol de clasificación (sin podar) mediante CART y utilizando como criterio la entropía. La condición de parada debe ser que los nodos hoja sean puros (todos los ejemplos sean de la misma clase). En cada nodo del árbol se debe indicar:

   - La variable y su valor umbral.
   - La entropía correspondiente.
   - En los nodos hoja, la clase del nodo y los ejemplos que pertenecen al mismo.

   Nota: este ejercicio debe hacerse sin utilizar ninguna función de scikit-learn.

2. Dado el problema de clasificación Blood Transfusion Service Center:

   a. La clase que implementa el algoritmo CART en problemas de clasificación en scikit-learn es *sklearn.tree.DecisionTreeClassifier*. Revisa los parámetros y métodos que tiene. En esta versión de scikit-learn aún no está implementado el método de podado del árbol.

   b. Divide los datos en entrenamiento (80%) y test (20%).

   c. Realiza la experimentación con *DecisionTreeClassifier* usando los valores por defecto de los parámetros, excepto para *criterion* que debe tomar el valor '*entropy*'. Además, utiliza como hiper-parámetro la variable *min_samples_split* (permitirá modificar el tamaño del árbol).

   i. Muestra la gráfica del error de entrenamiento con validación cruzada (5-CV) frente al valor del hiper-parámetro, y justifica la elección del valor más apropiado.

   ii. Muestra la gráfica del error de test frente al valor del hiper-parámetro, y valora si la gráfica del error de entrenamiento con validación cruzada ha hecho una buena estimación del error de test.

3. Repite el ejercicio 2 pero para el problema de regresión Energy Efficiency con la variable de salida *cooling load*. La clase que implementa el algoritmo CART en problemas de regresión en scikit-learn es *tree.DecisionTreeRegressor*.

4. ¿Crees que sería de interés aplicar un método de selección de variables (*Forward stepwise selection*, etc.) junto con el algoritmo CART?