Boletín 3: árboles

Para la realización de las prácticas correspondientes a este boletín se utilizará scikit-learn en el CESGA. Debes seleccionar **SEED_VALUE=1**.

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 son 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.

   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).

      Muestra la gráfica del error de entrenamiento con validación cruzada (5-CV) frente al valor del hiper-parámetro. ¿Cuál es el menor error de validación cruzada, su desviación estándar y el valor del hiper-parámetro para el que se consigue? ¿Cuál es el valor del hiper-parámetro si se aplicase la regla de una desviación estándar? En caso de que haya varios modelos con error mínimo, debe seleccionarse siempre el más simple.

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. ¿Cuál es el error de test para el valor del hiper-parámetro seleccionado por la validación cruzada?

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*.

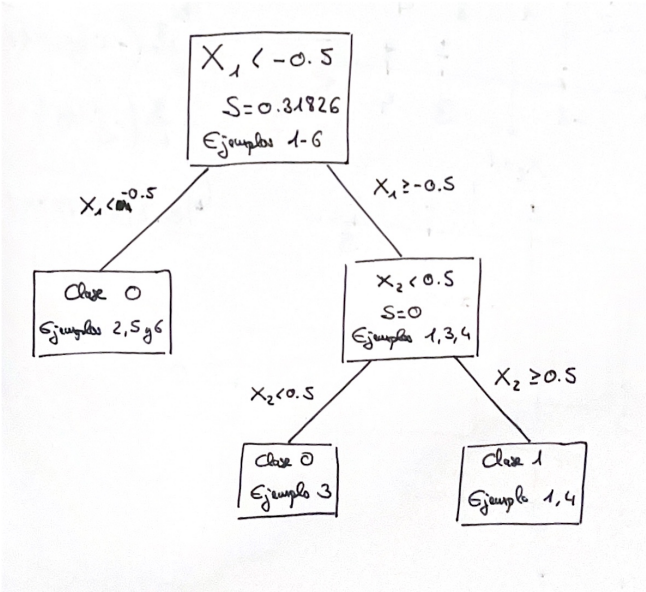
Entregable

Se debe entregar un único fichero comprimido con el nombre *PrimerApellido_SegundoApellido.zip* (también son válidos los formatos .rar y .7z), que contenga dos archivos:

- El primer archivo debe ser de tipo pdf, y contendrá exclusivamente las respuestas a los ejercicios (incluyendo las gráficas necesarias para justificar dichas respuestas). No se incluirá en este archivo ningún otro tipo de texto.

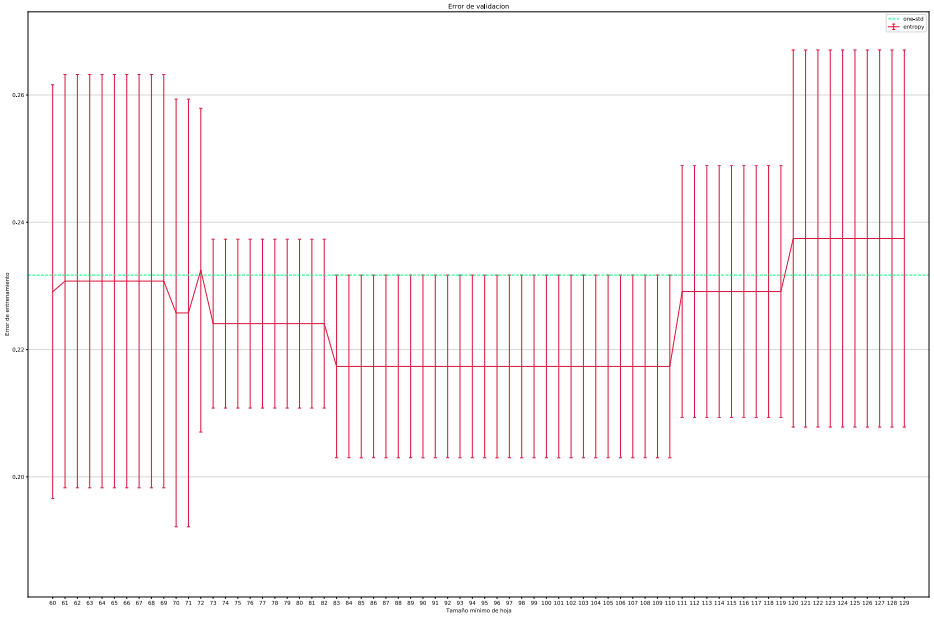
- El segundo archivo será de tipo ipynb, y permitirá reproducir toda la experimentación realizada en el boletín.

Ejercicio 1

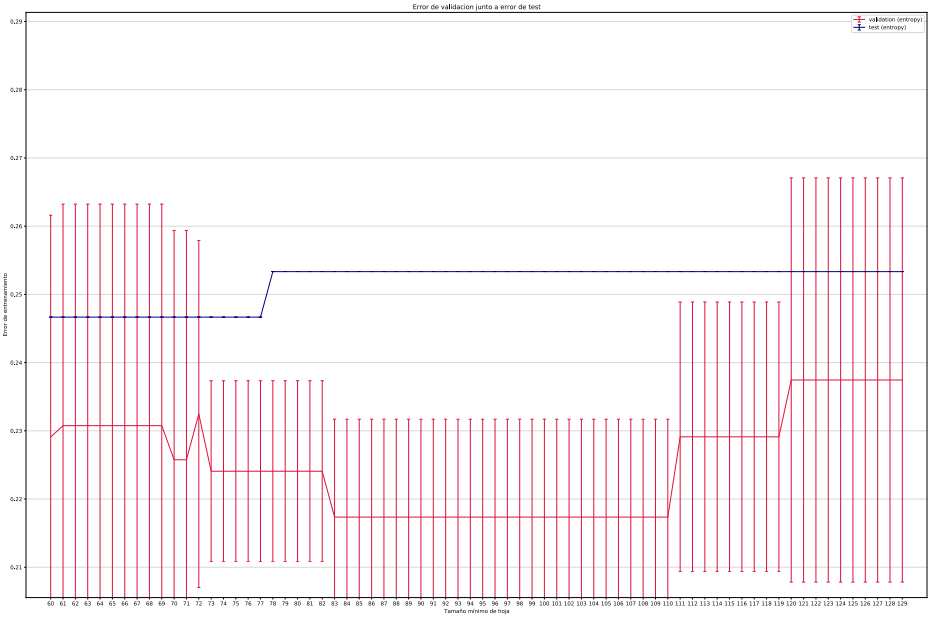


Ejercicio 2

- **Menor error de validación cruzada, su desviación estándar y valor del hiperparámetro:** $\Delta = 0{,}217353$, $\sigma = 0{,}014349$, $\text{hy} = 110$.
- **Con regla de una desviación estándar:** $\Delta = 0{,}229118$, $\sigma = 0{,}019768$, $\text{hy} = 119$.

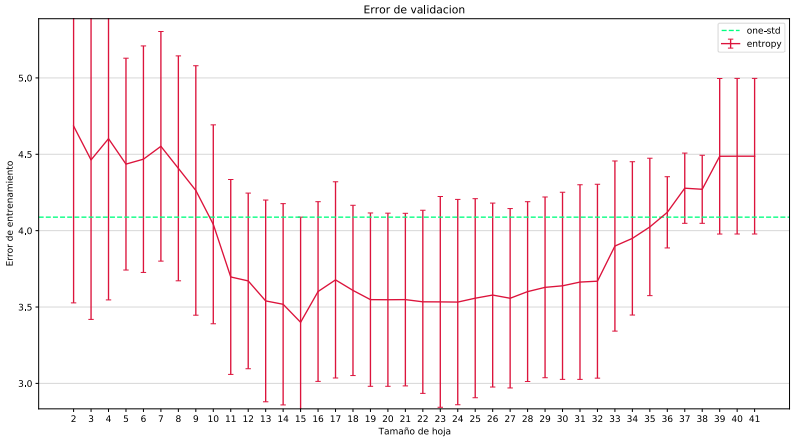


- **Error de test para el hiperparámetro de validación cruzada:** $\Delta = 0{,}253333$, $\text{hy} = 110$.

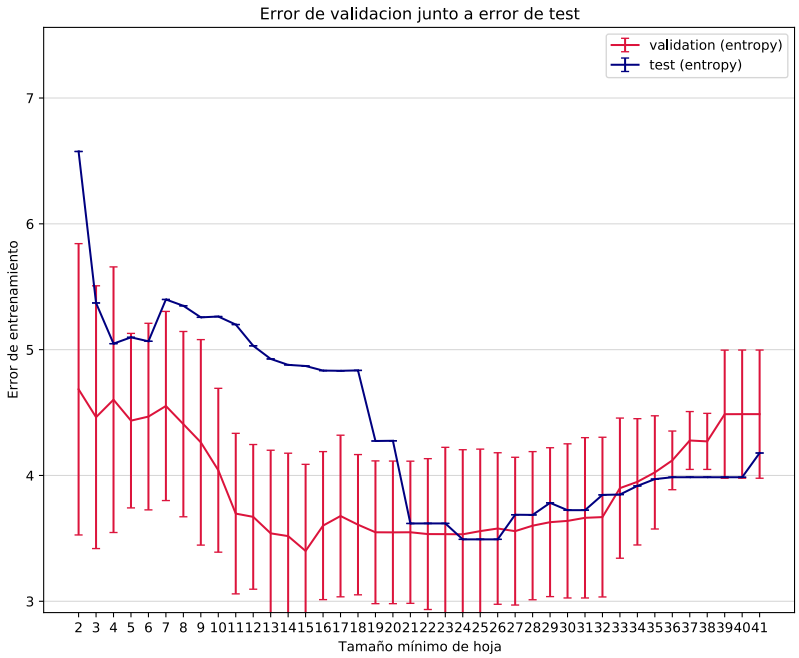


Ejercicio 3

- **Menor error de validación cruzada, su desviación estándar y valor del hiperparámetro:** MSE = 3,400035, σ = 0,68836, hy = 15.
- **Con regla de una desviación estándar:** MSE = 4,024361, σ = 0,449845, hy = 35.



- **Error de test para el hiperparámetro de validación cruzada:** MSE = 4,87007, hy = 15.


Error de validacion junto a error de test
validation (entropy)
test (entropy)
Error de entrenamiento
Tamaño mínimo de hoja