Tarea 1: Árboles de Decisión

Problema

Considere a «Predict students' dropout and academic success»[1], un juego de datos con 36 atributos y más de 4000 instancias. Tenga en cuenta que:

- hay atributos continuos que deberán ser preprocesados[2];
- el objetivo es el atributo *Target* y sus valores *enrolled* y *graduate* serán considerados como uno solo.

Se pide:

a) Implemente el algoritmo ID3 visto en el teórico agregándole los siguientes hiperparámetros:
   i) *min_samples_split*: cantidad mínima de ejemplos para generar un nuevo nodo; en caso de que no se llegue a la cantidad requerida, se debe formar una hoja.
   ii) *min_split_gain*: ganancia mínima requerida para partir por un atributo; si ningún atributo llega a ese valor, se debe formar una hoja.
b) Entrene y evalúe los resultados de su implementación utilizando el dataset preprocesado.
c) Discuta cómo afecta la variación de esos hiperparámetros a los modelos obtenidos.
d) Corra los algoritmos de scikit-learn *DecisionTreeClassifier*[3] y *RandomForestClassifier*[4] y compare los resultados.

Se podrá utilizar pandas y scikit-learn para la carga del dataset, su preprocesamiento y la generación de archivos de entrenamiento, testeo, etc.

Entregables

- Informe con las pruebas realizadas y los resultados obtenidos.
- El informe a entregar debe ser un Jupyter Notebook.
- Código escrito para resolver el problema.

Fecha límite de entrega

Miércoles 30 de agosto (inclusive).

[1] https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success
[2] https://scikit-learn.org/stable/modules/preprocessing.html
[3] https://scikit-learn.org/stable/modules/tree.html
[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html