Tarea 1: Árboles de Decisión

Objetivo

Este laboratorio tiene por objetivos: a) implementar el algoritmo ID3 con la extensión para atributos numéricos; b) el uso de scikit-learn para el preprocesamiento de datos y la creación de modelos basados en árboles de decisión; y c) la evaluación de los modelos generados.

Problema

Considere al conjunto de datos «AIDS Clinical Trials Group Study 175»[1], con más de 2000 instancias y 23 atributos:

- Implemente el algoritmo ID3 visto en el teórico, agregando el siguiente hiperparámetro:

  max_range_split: cantidad máxima de rangos en los que se puede partir un atributo número.

- Entrene y evalúe los resultados de su implementación con valores de *max_range_split* de 2 y 3.
- Preprocese los valores numéricos justificando el proceso realizado.
- Entrene y evalúe los resultados de su implementación con el dataset preprocesado.
- Entrene modelos con los algoritmos de scikit-learn *DecisionTreeClassifier*[2] y *RandomForestClassifier*[3].
- Compare los resultados de todos los modelos generados.

Se podrá utilizar pandas y scikit-learn para la carga del dataset, su preprocesamiento[4] 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 4 de setiembre (inclusive).

1https://archive.ics.uci.edu/dataset/890/aids+clinical+trials+group+study+175
2https://scikit-learn.org/stable/modules/tree.html
3https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
4https://scikit-learn.org/stable/modules/preprocessing.html