Introducción a la Ciencia de los Datos

Tarea 9

Nota: entregar las respuestas el próximo miércoles al correo de Daniela ([daniela.pena@alumnos.cide.edu](mailto:daniela.pena@alumnos.cide.edu)) en un archivo: nombre\_apellido\_tarea9. py

(i) A partir de la base de datos Iris que está disponible en sklearn construye cuatro árboles de decisión con niveles de profundidad máximos de 1, 2, 3 y ninguno. Utiliza 33% de los datos para el conjunto de prueba

(ii) Imprime pruebas de exactitud con los conjuntos de entrenamiento y de prueba. ¿Qué modelo se te hace el más adecuado?, ¿Crees que haya overfitting? ¿Crees que el modelo con profundidad = 1 es bueno? ¿Por qué los resultados con tres niveles y sin especificar profundidad máxima son iguales? ¿Qué significa que el valor de Gini de cada etapa se reduzca?

(iii) ¿Calcula la importancia relativa de los atributos?

(iv) Para la misma base de datos construye árboles de decisión con profundidad = 3, pero ahora en vez de utilizar los cuatro atributos usa por separado todas las combinaciones posibles de parejas.

(v) Imprime pruebas de exactitud con los conjuntos de entrenamiento y de prueba.

(vi) Crees que tus resultados son consistentes con el resultado de importancia relativa de los atributos obtenido en el caso en que se usan los cuatro atributos para el análisis.