Trabajo Práctico Machine Learning

Introducción a la Inteligencia Artificial

Mellino, Natalia Farizano, Juan Ignacio

Introducción al dataset

Trabajaremos sobre el conjunto de datos *yeast* donde entrenaremos a nuestros modelos para que clasifiquen distintos tipos de levadura.

Este dataset consta de 9 atributos, el primero es simplemente un nombre de secuencia y por lo tanto se ignora ya que no tiene relevancia. Los restantes atributos son: MCG, GVH, ALM, MIT, ERL, POX, VAC, y NUC.

Cada instancia será clasificada en alguna de las siguientes 10 clases, donde originalmente en los datos dados se distribuyen de la siguiente manera:

Clase	Cantidad
CYT	463
NUC	429
MIT	244
ME3	163
ME2	51
ME1	44
EXC	37
VAC	30
POX	20
ERL	5

Entrenamiento del modelo

Al entrenar el modelo, utilizamos el método de *k-fold cross validation*, donde elegimos un k igual 5, para obtener nuestros conjuntos de entrenamiento y validación.

Una vez obtenidos estos conjuntos, procedemos a crear nuestros árboles de decisión utilizando la librería *rpart* y su función homónima.

En primera instancia, como criterio elegido para dividir los nodos utilizamos el método de Ganancia de Información y luego, repetimos el mismo proceso pero con el la medida de Impureza de Gini. En la siguiente sección podemos observar un análisis más detallado sobre ambos modelos obtenidos y una comparación entre estos dos métodos.

Resultados

Utilizando Ganancia de Información

Accuracy

Precision y Recall

Pruning

Utilizando el método de Gini

Accuracy

Precision y Recall

Pruning

Comparación

Conclusiones