Clasificadores simples de documentos reales.

Los conjuntos de datos ReutersCorn-train.arff y ReutersGrain-train.arff son conjuntos de datos de entrenamiento derivados de colecciones de artículos que se utilizan como referencia para evaluar clasificadores de documentos. ReutersCorn-test.arff y ReutersGrain-test.arff son sus correspondientes conjuntos de prueba.

Los documentos en los conjuntos Corn y Grain son los mismos. Solo difieren en las etiquetas de clase. En el primer conjunto de datos, los artículos relacionados con Corn tienen el valor de clase 1 y los restantes 0. El objetivo es construir un clasificador que identifique artículos relacionados con Corn. En el segundo conjunto de datos las etiquetas se elaboran para los artículos relacionados con Grain.

Ejercicio 1: Crear clasificadores para los dos conjuntos de datos, con *FilteredClassifier,* aplicando *StringToWordVector* con J.48 y NBMultinomial, evaluándolos sobre el correspondiente conjunto de test.

1-a. ¿Qué porcentaje de clasificación correcta se obtiene en los cuatro escenarios?
1-b. En base a las tasas de error, ¿qué clasificador elegiría para cada conjunto de datos?

Ejercicio 2: En la tabla *Detailed Accuracy by Class* se calcula el área bajo la curva ROC, AUC. Weka elabora la curva ROC de cada clase, que denomina Threshold curve. Se visualiza pinchando sobre el último experimento realizado con el botón derecho.

2-a. Comparar las curvas ROC de ambos clasificadores para la clase de interés en cada conjunto de datos.
2-b. En base a las curvas ROC, ¿qué clasificador elegiría para cada conjunto de datos?

Ejercicio 3: En la clasificación de documentos se utilizan otras métricas para evaluar los clasificadores, como *precisión* y recall, o la *medida-F.* Todas ellas se elaboran a partir de TP, FP, TN, FN. Sus valores se incluyen en la tabla *Detailed Accuracy by Class*.

En base a sus definiciones, ¿cuáles son los mejores posibles valores para estas métricas? Describir en que circunstancias se obtiene estos mejores valores.