PRÁCTICA 6:
CLASIFICADORES SIMPLES DE DOCUMENTOS REALES.

MINERÍA DE DATOS
PATRICIA AGUADO LABRADOR

EJERCICIO 1

A)

PORCENTAJES DE ACIERTO SOBRE LOS CONJUNTOS DE TEST		
	J48	**NBMultinomial**
Corn	97,351 %	93,709 %
Grain	96,358 %	90,729 %

Utilizando J48 conseguimos tasas de acierto cercanas al 100%, se comporta mejor que NBMultinominal en estos conjuntos de datos.

B)

TASAS DE ERROR SOBRE LOS CONJUNTOS DE TEST		
	J48	**NBMultinomial**
Corn	0,027	0,063
Grain	0,036	0,093

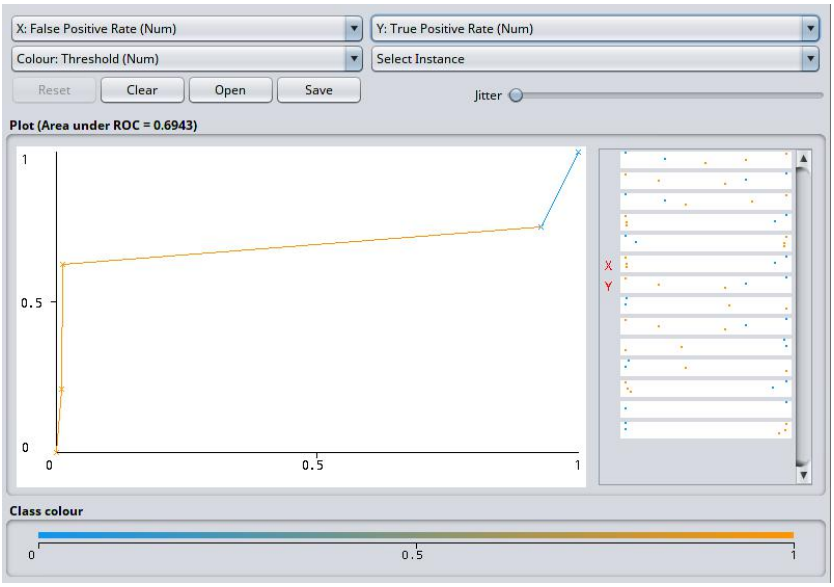
Basándome en la tasas de error para ambos conjuntos de datos eligiría el algoritmo J48 porque obtenemos tasas de error mucho menores, aproximadamente la mitad de las tasas de error del otro algoritmo.

EJERCICIO 2

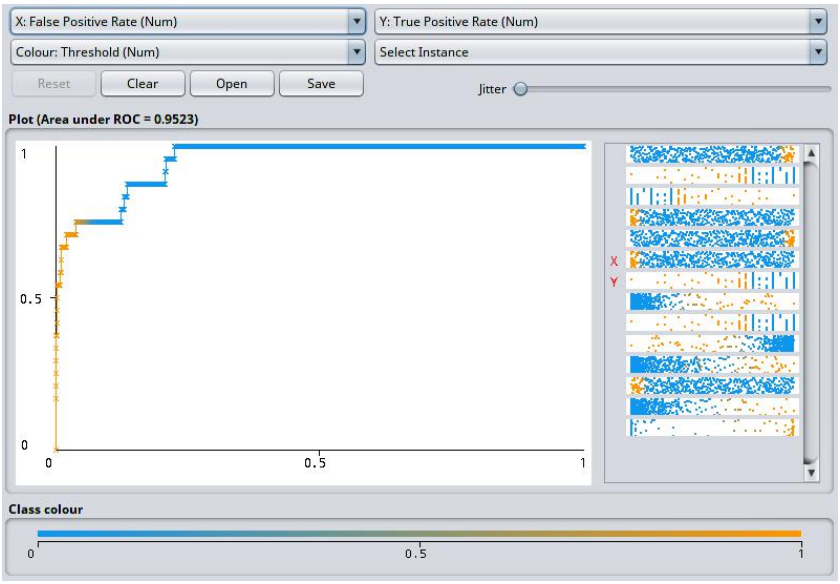
A)

CORN	AUC (Área bajo la curva ROC)
J48	0,694
NBMultinomial	0,952

J48



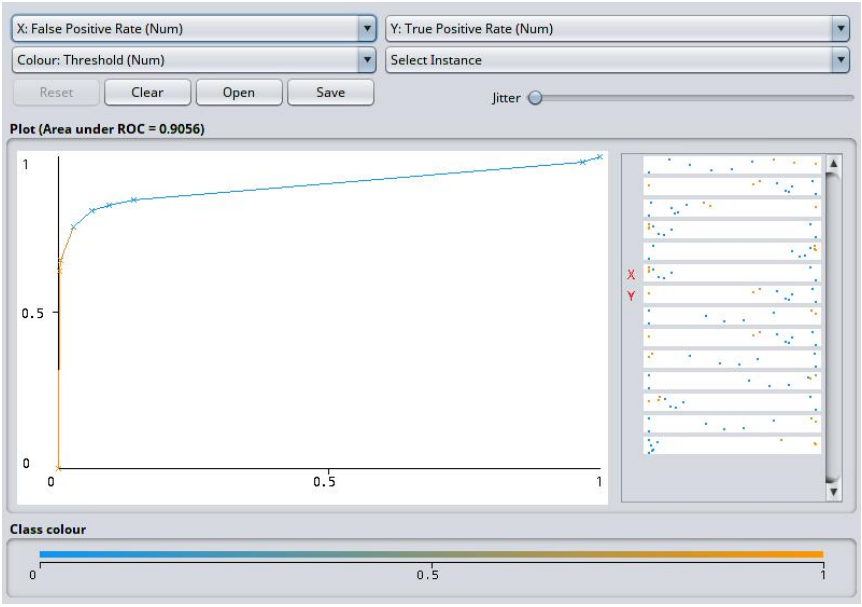
NBMultinomial



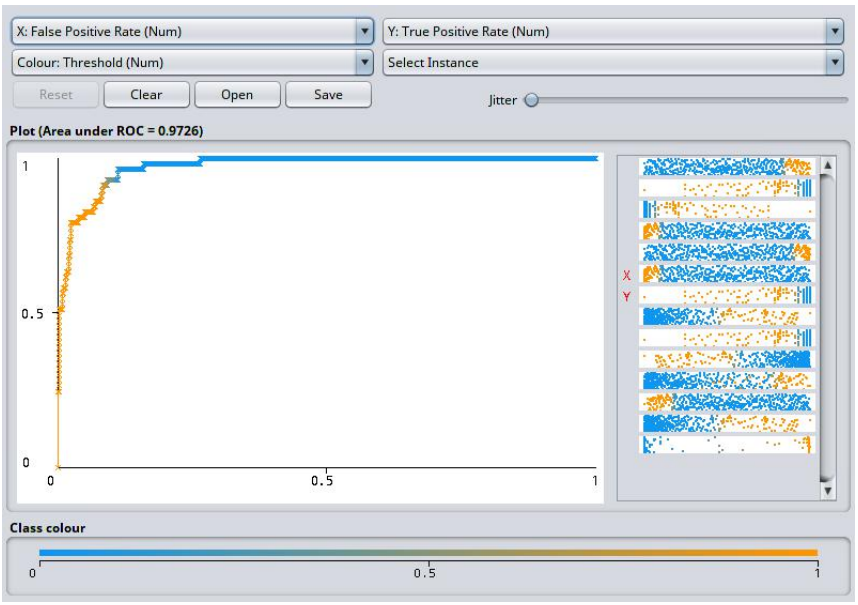
Para el conjunto de datos relacionado con Corn el algoritmo que mejor funciona es Naive Bayes Multinominal ya que al comparar el área bajo la curva, este algoritmo sale ganando.

GRAIN	AUC (Área bajo la curva ROC)
J48	0,906
NBMultinomial	0,973

J48



NBMultinomial



En este otro caso ocurre lo mismo, aunque la diferencia entre las áreas bajo la curva sea menor, Naive Bayes obtiene un resultado mayor.

EJERCICIO 3

Precission: porcentaje de instancias recuperadas que son relevantes (TP/(TP+FP)). Si este valor es próximo a 1 entonces se entiende que todos los documentos que se recuperan son relevantes. Cuanto menor sea el número de falsos positivos detectado mayor será la precisión.

Recall: porcentaje de instancias relevantes que han sido recuperadas (tp = TP/P). Si este valor es próximo a 1 entonces se entiende que se recuperan todos los documentos que son relevantes. El mejor valor para esta métrica es que el valor de los ciertos positivos sea próximo al total de positivos clasificados.

F-meassure: medida de precisión de un test que se obtiene ponderando la precisión y el recall, en el caso de esta formula estamos dando la misma importancia a ambos valores ((2*recall*precisión)/(recall+precisión)). Cuanto más mayores sean los valores de precisión y recall más preciso será el test que se realice.

Conjunto de datos Corn			
	Precission	**Recall**	**F-meassure**
J48	0,682	0,625	0,652
NBMultinomial	0,36	0,75	0,486

Conjunto de datos Grain			
	Precission	**Recall**	**F-meassure**
J48	0,966	0,995	0,98
NBMultinomial	0,99	0,907	0,947

Viendo estos resultados, para el conjunto de datos Corn eligiría el método J48 ya que tiene mejores resultados a pesar de que NBMultinomial tenga mayor Recall. En el caso del conjunto de datos Grain elegiría también J48 a pesar de que ocurre lo mismo que para Corn, pero en este caso con la precisión.