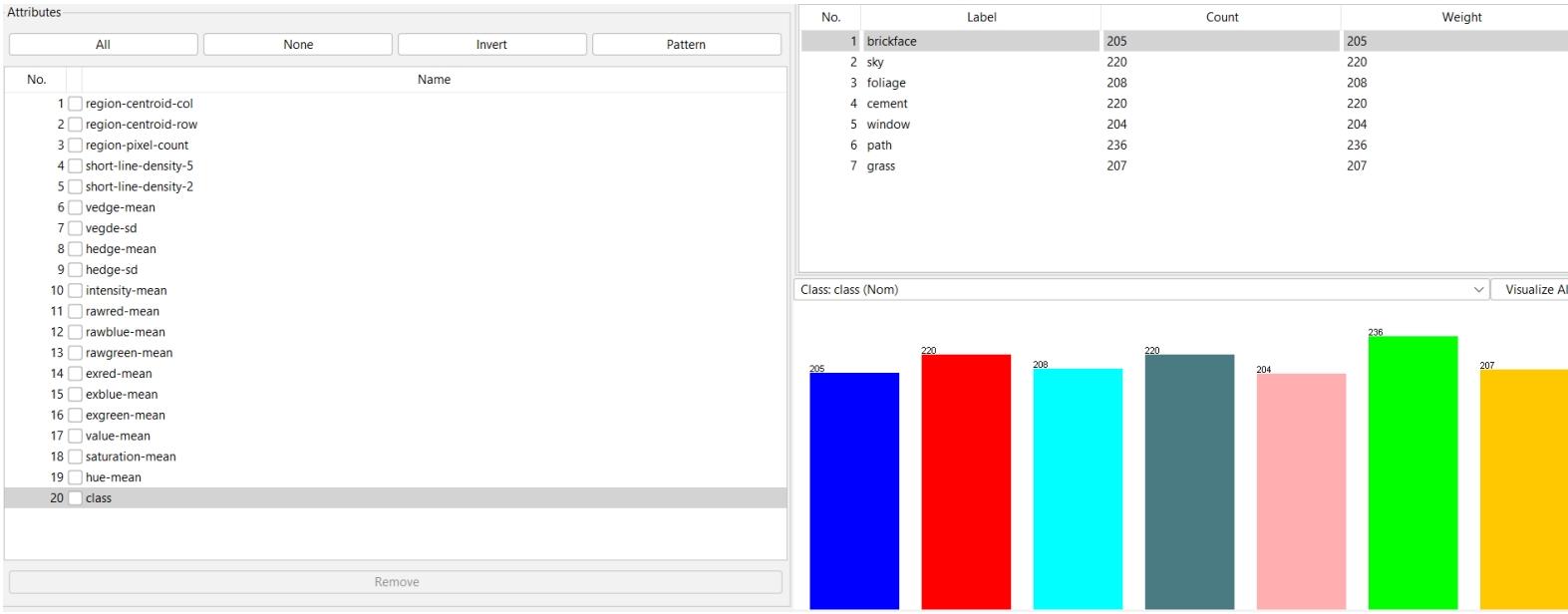
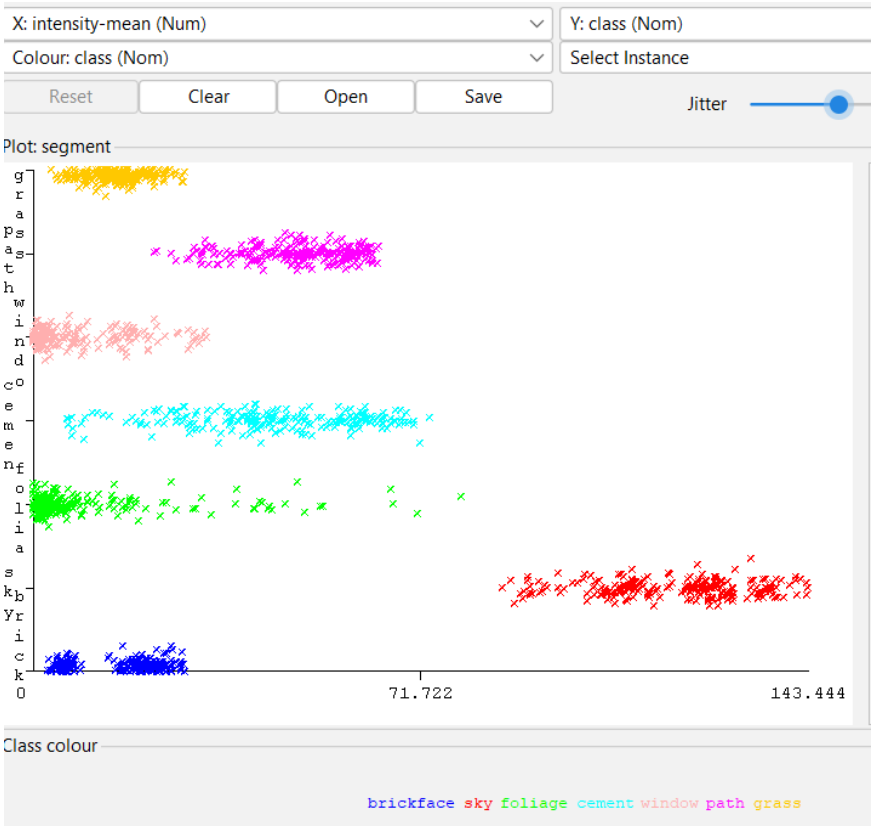
Al cargar el .arff visualizamos lo siguiente:



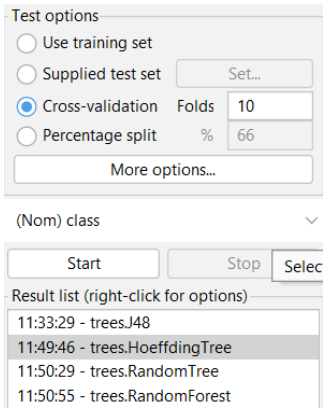
Los atributos implican que estamos muy probablemente analizando una o varias imágenes debido a que están todos relacionados con colores. Debido a que varios atributos son una media, es probable que estos datos correspondan a áreas o porciones de una o varias imágenes, lo que debería reducir los outliers.

En esta situación, analizando los datos mirando la clase en función de atributos, podemos empezar a hacer observaciones:



Por ejemplo, cuando observamos el atributo intensity-mean, observamos que es muy fácil determinar si una porción de imagen tiene la clase “sky”, simplemente hay que mirar si intensity-mean > 86. Este tipo de clasificación de datos corresponde a un árbol binario, así que probemos usar algunos de estos modelos con weka.

Probando los árboles de tipo J48, Hoeffding, Random Tree y el modelo Random Forest, con los siguientes parámetros,



obtengo los siguientes resultados:
El árbol Hoeffding clasifica correctamente el 80.5% de los valores, este árbol es mucho menos eficiente que los tres otros con creces, sin embargo, si miramos la exactitud(accuracy) por clase, vemos lo siguiente:

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.951	0.053	0.741	0.951	0.833	0.812	0.980	0.898	brickface
0.986	0.000	1.000	0.986	0.993	0.992	0.998	0.994	sky
0.207	0.014	0.705	0.207	0.320	0.337	0.943	0.649	foliage
0.850	0.028	0.839	0.850	0.844	0.817	0.960	0.864	cement
0.701	0.127	0.464	0.701	0.559	0.487	0.887	0.509	window
0.936	0.004	0.978	0.936	0.957	0.949	0.998	0.976	path
0.976	0.000	1.000	0.976	0.988	0.986	0.999	0.994	grass
0.805	0.031	0.824	0.805	0.791	0.775	0.967	0.845	

Obtiene unos datos muy decentes para la mayoría de tipos, sin embargo, es incapaz de detectar la clase foliage con solo un 20% de detección. Por lo tanto, lo más sensato es descartar el modelo. Este modelo es el segundo más rápido en ser construido con un tiempo de 60 milisegundos

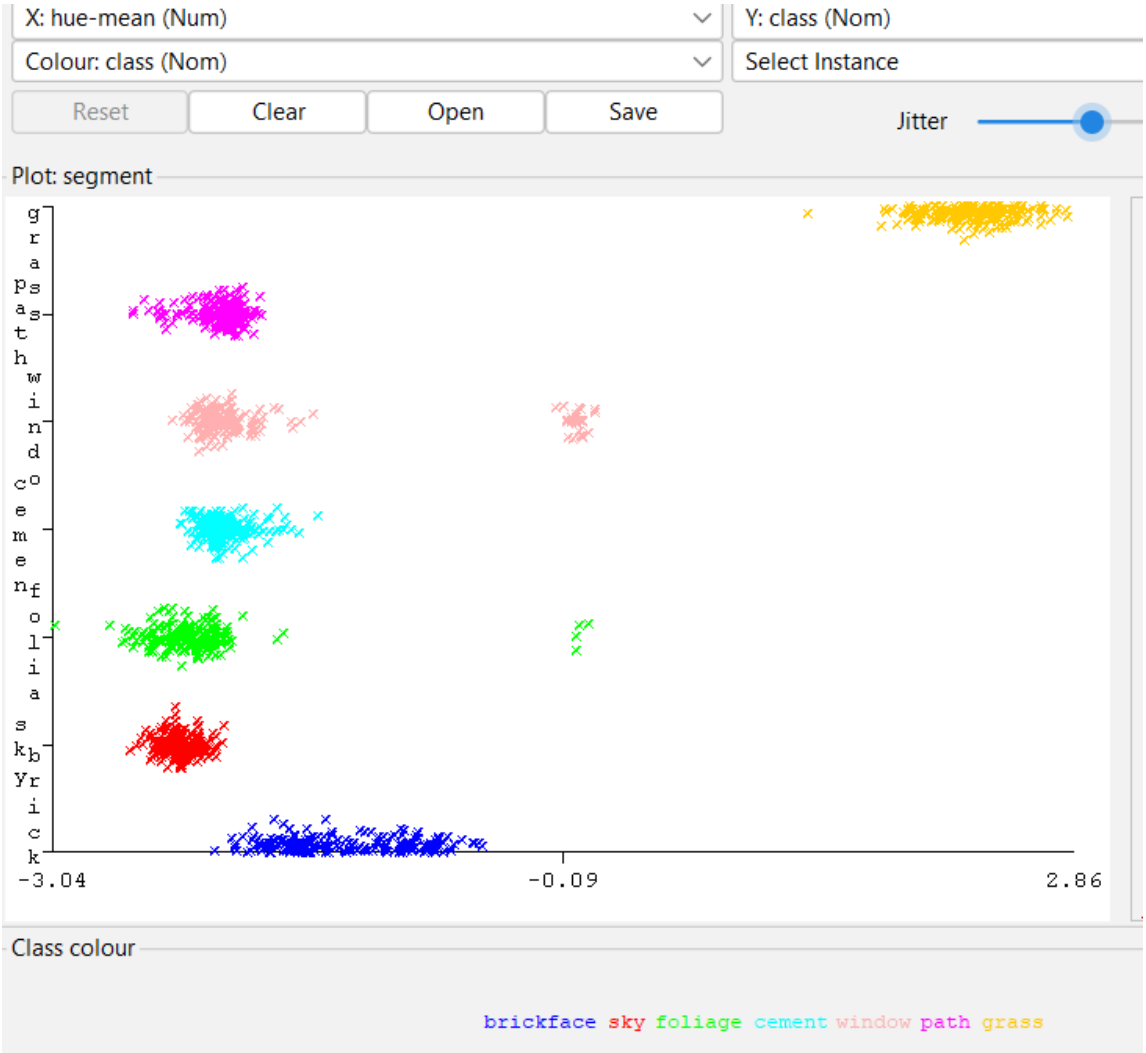
El modelo random tree es tercero en tasa de exactitud, con un 94.7%, este resultado es muy satisfactorio ya que este modelo es el más rápido de construir con un tiempo de solo 20 milisegundos. Es 15 veces más rápido que RandomForest y 7 veces más rápido que J48 por lo tanto un resultado excelente.

El árbol J48 es el segundo con mejor tasa de exactitud, clasificando correctamente el 95.7% de los datos, este porcentaje es satisfactorio. al estar por encima de 95%. Sin embargo está un poco en medio, al tardar 140 milisegundos, es muy lento comparado con random tree y tiene solo un punto porcentual más en tasa de exactitud, por tanto, no sería el mejor modelo.

El modelo Random Forest es el más caro en tiempo, usando 300 milisegundos, sin embargo, tiene la mayor tasa de exactitud, con un 97.87% de clasificaciones correctas, esta tasa es extremadamente satisfactoria. El hecho de haber usado solo 10 folds de cross-validation nos permite evitar el overfitting. Y descomponiendo la exactitud por clase, podemos observar que el modelo llega a tener una tasa del 100% para las clases sky y grass:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.990	0.001	0.995	0.990	0.993	0.992	1.000	1.000	brickface
	1.000	0.001	0.995	1.000	0.998	0.997	1.000	1.000	sky
	0.966	0.009	0.948	0.966	0.957	0.950	0.998	0.988	foliage
	0.968	0.006	0.964	0.968	0.966	0.960	0.998	0.993	cement
	0.926	0.008	0.950	0.926	0.938	0.928	0.997	0.983	window
	0.996	0.001	0.996	0.996	0.996	0.995	1.000	1.000	path
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	grass
Weighted Avg.	0.979	0.003	0.979	0.979	0.979	0.975	0.999	0.995	

Esto era de esperar ya que cómo hemos visto antes, es muy fácil determinar los elementos de clase sky usando el atributo intensity-mean.



Observando la clase en función del atributo hue-min, podemos ver que también se puede determinar fácilmente si un elemento es de tipo grass o no. Lo que explica la alta exactitud en determinar este atributo.