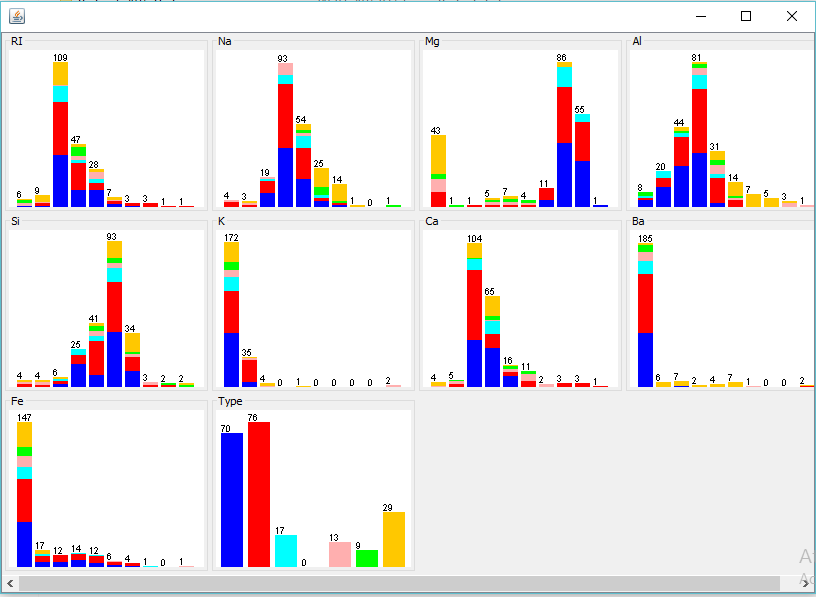
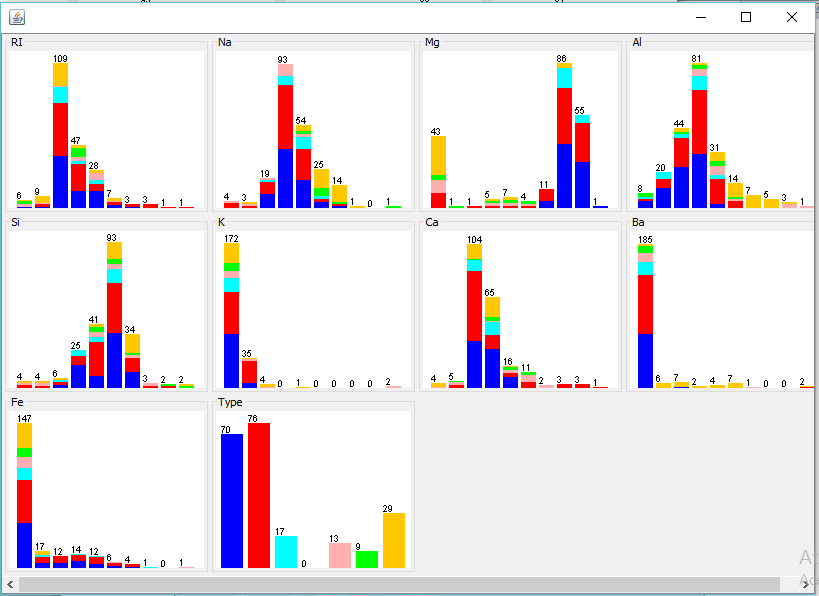
**Resposta da Tarefa 3:**

Nesta tarefa aplicaremos outra operação de pré-processamento que é a discretização de dados numéricos. Esta é uma operação utilizada nos casos em que temos conjuntos de dados com atributos numéricos e pretendemos converter os referidos valores à uma escala nominal.

Existem dois tipos de discretização: *não supervisionada*, não tem em conta as classes das instâncias, e *supervisionada*, tem em conta as classes das instâncias ao criar os intervalos. O principal método de *discretização não supervisionada* existente no Weka é *Weka.filters.unsupervised.attribute.Discretize*. O mesmo implementa duas variantes: *discretização com intervalos igual largura* (seleccionado por defeito) e com *intervalos de igual frequência*.

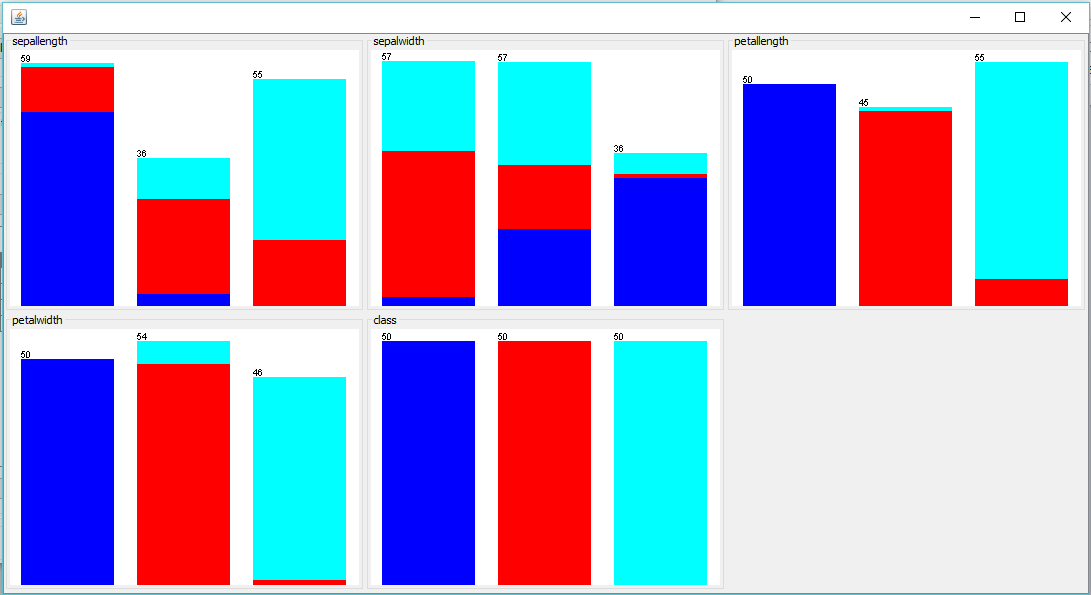
Carregamos o conjunto de dados *glass.arff* no explorador.

Aplicamos o método de *discretização não supervisionada* nos dois modos mencionados. Para tal no separador *Preprocess*, fez-se click no botão *Choose* e seleccionamos *filters > unsupervised > Discretize*. Obtivemos o seguinte resultado:

Para mudar de modo, fez-se click com o botão direito do rato na barra de edição diante do botão *Choose* e seleccionamos *Show properties* no menú de contexto. Na caixa de diálogo alteramos o valor da propriedade *useEqualFrequency de “False” para “True”*. Obtivemos o seguinte resultado:

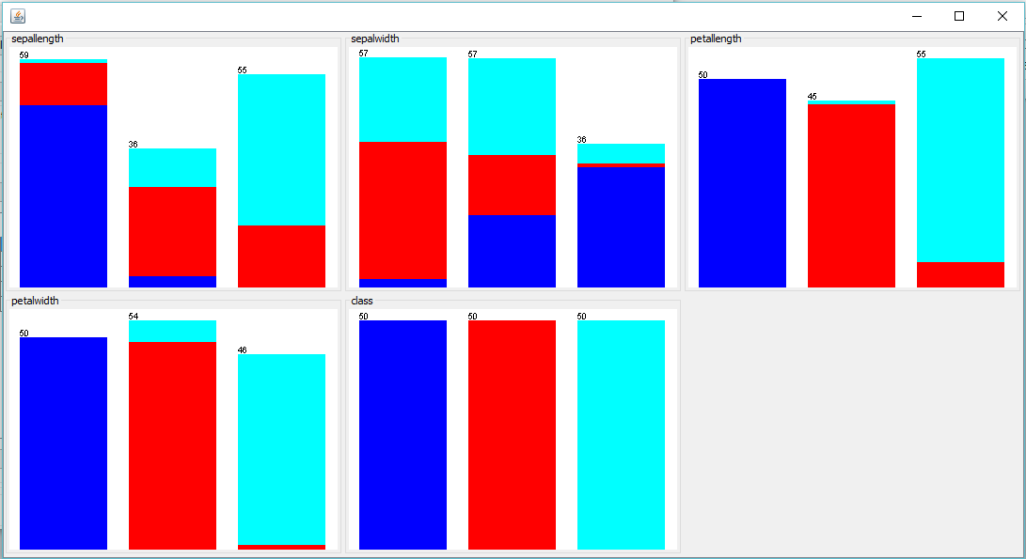
Observou-se o seguinte ao comparar os histogramas obtidos: Os histogramas correspondentes à *discretização com intervalos de igual frequência* **não** são bastante enviesados para **todos** os atributos, porque ao altermos o valor do atributo “*useEqualFrequency*” os atributos são experimentados na mesma frequência logo não há alterações significativa.

A principal técnica de *discretização supervisionada* de atributos numéricos no Weka é *weka.filters.supervised.attribute.Discretize*.

Carregamos o conjunto de dados *iris.arff*, aplicando a *discretização supervisionada* e observamos os histogramas obtidos. A *discretização supervisionada* criamos intervalos nos quais a distribuição das classes é consistente, ainda que a distribuição das instâncias varie de um intervalo para outro. Otivemos os seguintes gráficos:

Com base nos histogramas obtidos, o atributo discretizados que possui maior poder preditivo é o **sepalwidth.**

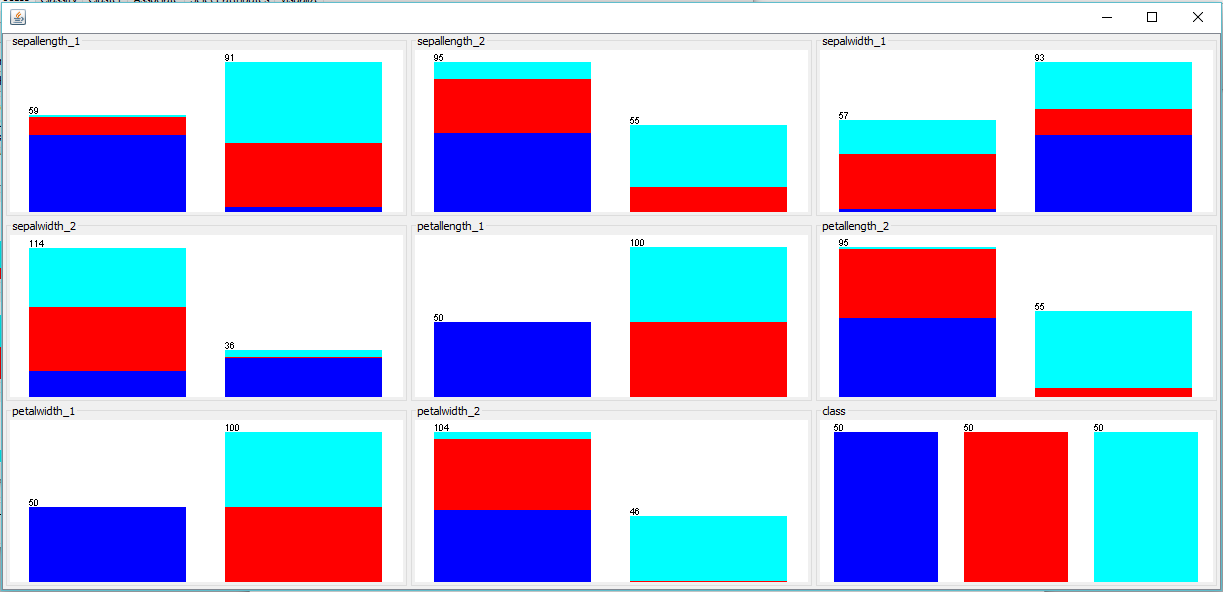
Carreguando novamente o conjunto *glass.arff* e aplicando a *discretização supervisionada*.



Para alguns atributos existe uma única barra no histograma isto significa que no referido atributo as instâncias possuem a mesma classififcação.

Os atributos discretizados são normalmente codificados como atributos nominais, com um único valor para cada rango. Ambos os filtros permitem a criação de atributos com valores binários em vez de multivalorados, activando a opção *makeBinary*, na caixa de diálogo de propriedades.

Escolhemos o conjunto de dados *“iris.arff”* e utlizando um dos métodos e usando-o para criar atributos binários. Comparando os resultados com as saídas produzidas quando a opção *makeBinary* é falsa. Observamos que os atributos binários são aqueles que possuem apenas duas correncias e que se o conjunto de dados estiver discretizado quase que não haverá diferencia quanto a opção *makeBinary* true ou false. Observou-se um único histograma para o conjunto de dados em questão.



Examinemos agora o efeito da discretização ao construir uma árvore de decisão com o algoritmo *C4.5* (designado *J48* no Weka) com os dados do conjunto *hionosphere.arff.* Este conjunto de dados contém informação sobre sinais de radar retornados da ionosfera. Consideram-se “Boas” instâncias aquelas que mostram evidências da existência de algum tipo de estrutura na ionosfera, enquanto que os sinais “Maus” são aqueles que atravessam directamente a ionosfera. Para mais informações, veja os comentários no ficheiro ARFF.

Iniciamos com a *discretização não supervisionada*.

Para o algoritmo *C4.5* (*J48*), *weka.classifiers.trees.J48*, comparamos a taxas de erro da validação cruzada e o tamanho das árvores geradas pelos (1) dados em bruto, (2) dados discretizados usando o método não supervisionado de discretização no modo por defeito e (3) dados discretizados pelo mesmo método com atributos binários.

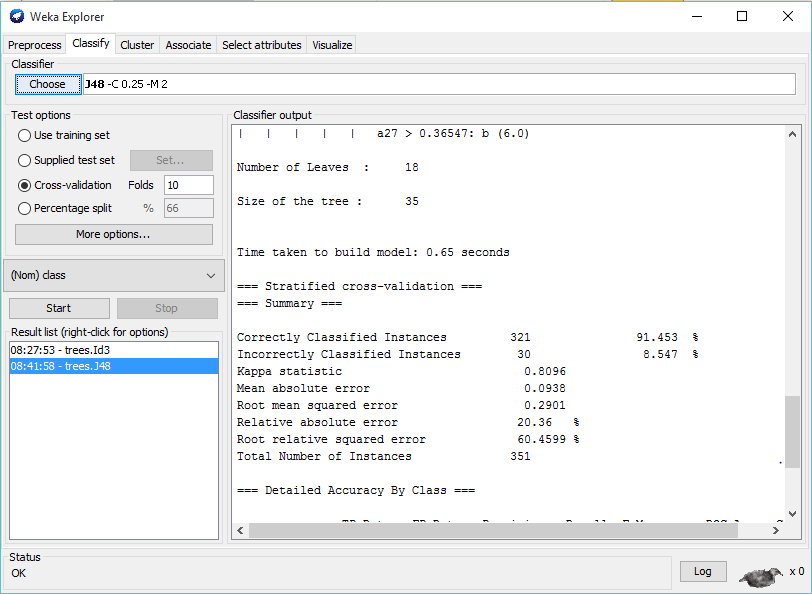
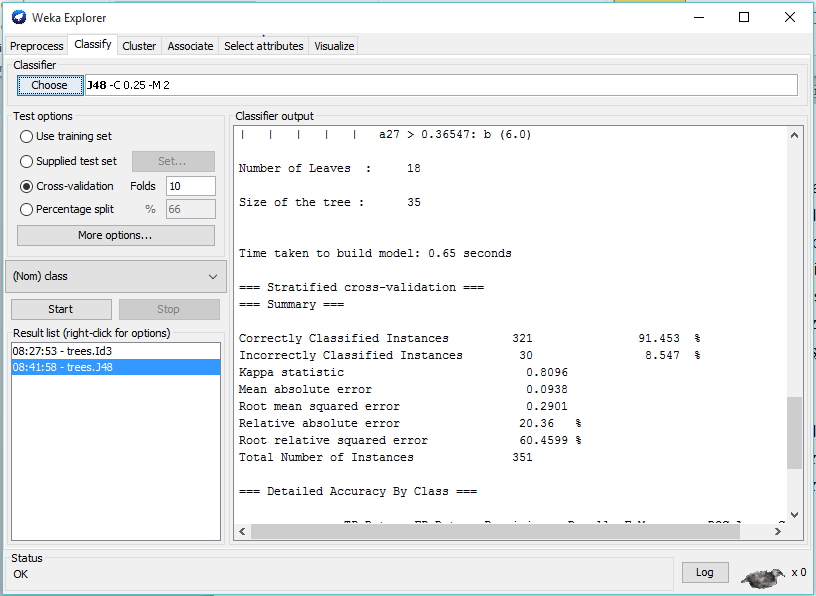


Imagem acima refletindo a taxas de erro da validação cruzada e o tamanho das árvores geradas pelos dados em bruto.



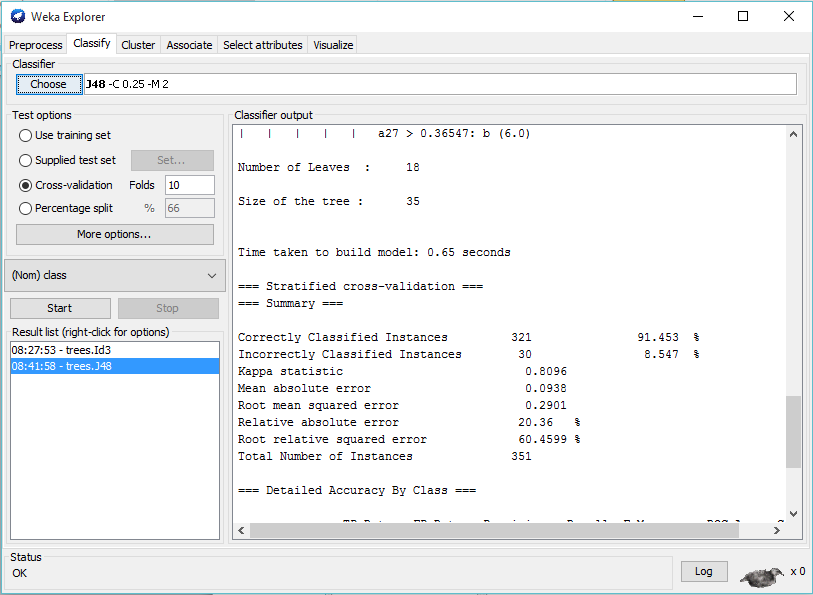
Imagem acima refletindo a taxas de erro da validação cruzada e o tamanho das árvores geradas pelos dados discretizados usando o método não supervisionado de discretização no modo por defeito.

Imagem acima refletindo a taxas de erro da validação cruzada e o tamanho das árvores geradas pelos dados discretizados usando dados discretizados pelo mesmo método com atributos binários.

Vejamos agora a *discretização supervisionada*. Neste caso, se utilizamos a totalidade dos dados para a discretização os resultados da validação cruzada são demasiadamente optimistas. O procedimento correcto é costruir o filtro utilizando apenas os dados de treino e avaliar o classificador sobre os dados de teste, utilizando os intervalos de discretização determinados com os dados de treino. Para tal se utiliza um *metaclassificador* incluído no Weka, *FilteredClassifier* (*weka.classifiers.meta. FilteredClassifier*).

Utilizando *FilteredClassifier* e *J48*, comparando a taxa de erro da validação cruzada e o tamanho das árvores geradas pela (4) *discretização supervisionada*, no modo definido por defeito, e (5) *discretização supervisionada* com atributos binários.

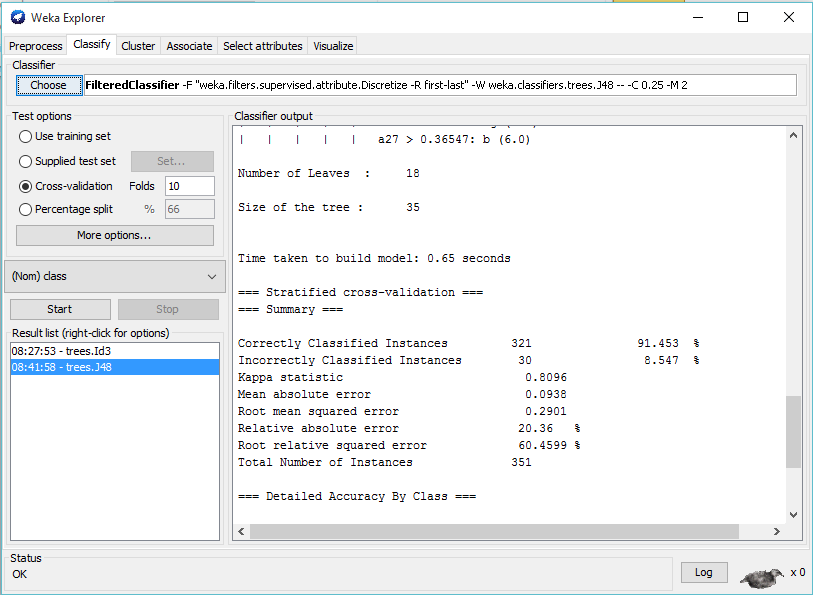


Imagem acima refletindo a taxa de erro da validação cruzada e o tamanho das árvores geradas pela *discretização supervisionada*, no modo definido por defeito.

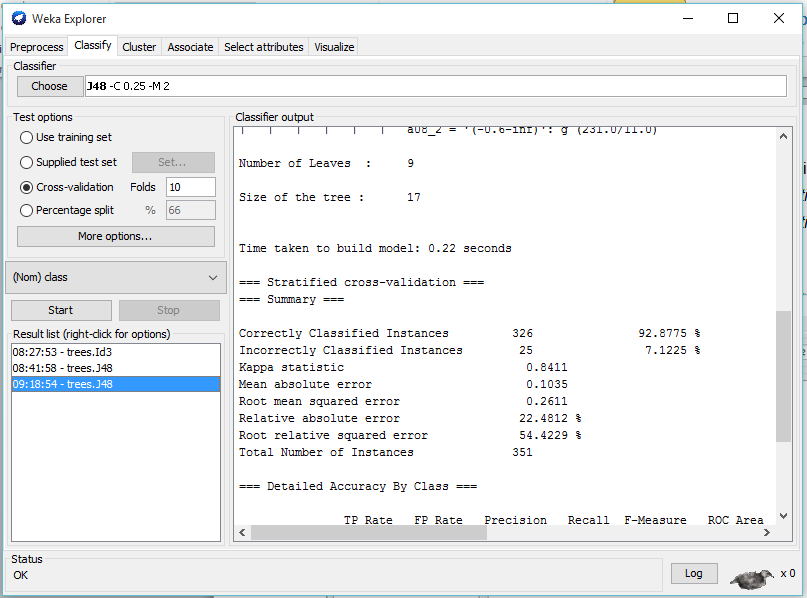


Imagem acima refletindo a taxa de erro da validação cruzada e o tamanho das árvores geradas pela *discretização supervisionada* com atributos binários.

Em função da nossa análise constatou-se que nesse caso não houve diferença entre a árvore de decisão gerada apartir dos dados brutos e a árvore de decisão gerada por dados discretizados.