

Universidade Federal de Uberlândia

Aula Prática - 08

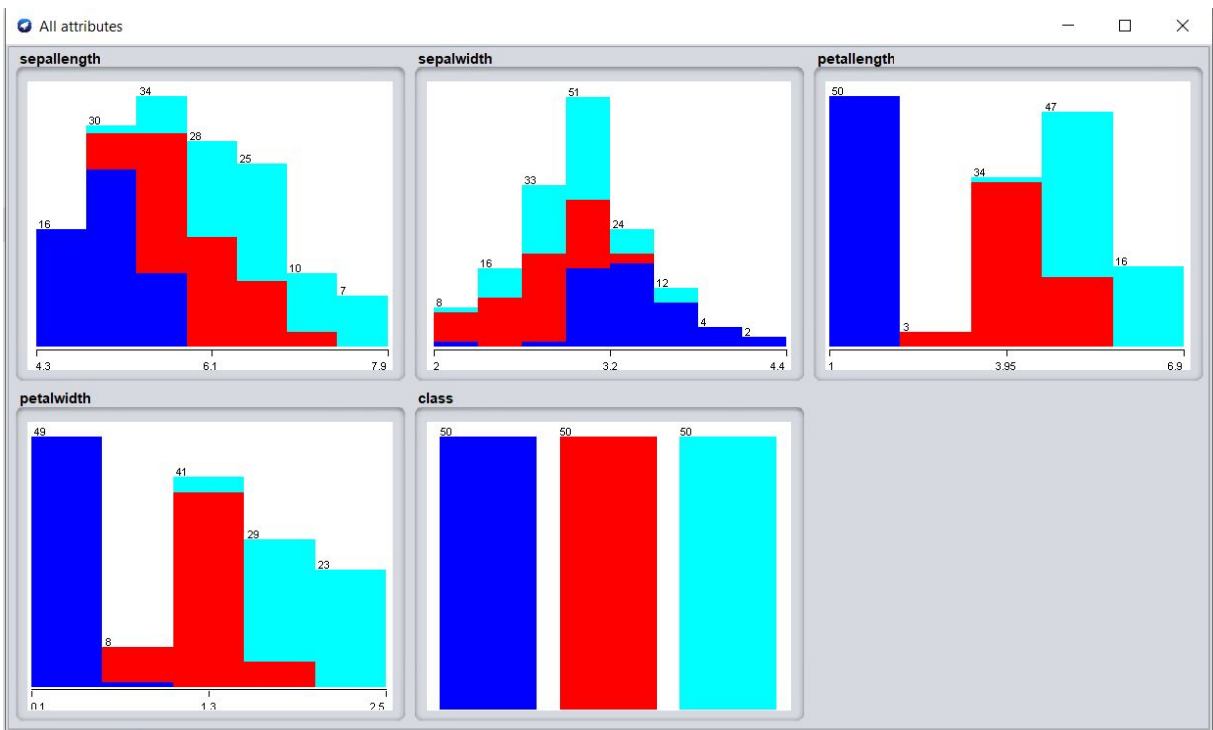
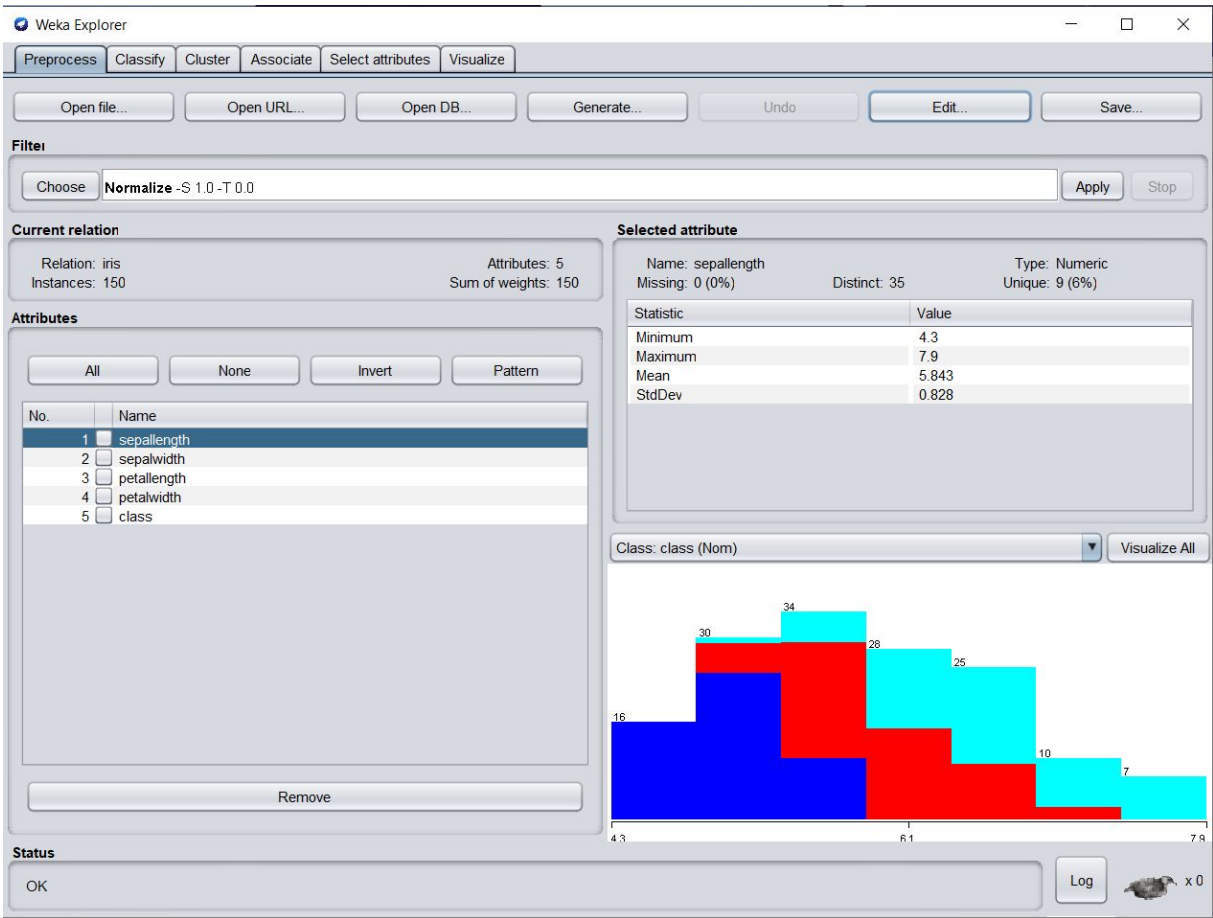
Pedro Henrique Faria Teixeira - 11621BCC025

João Daniel de Aquino Rufino - 11621BCC033

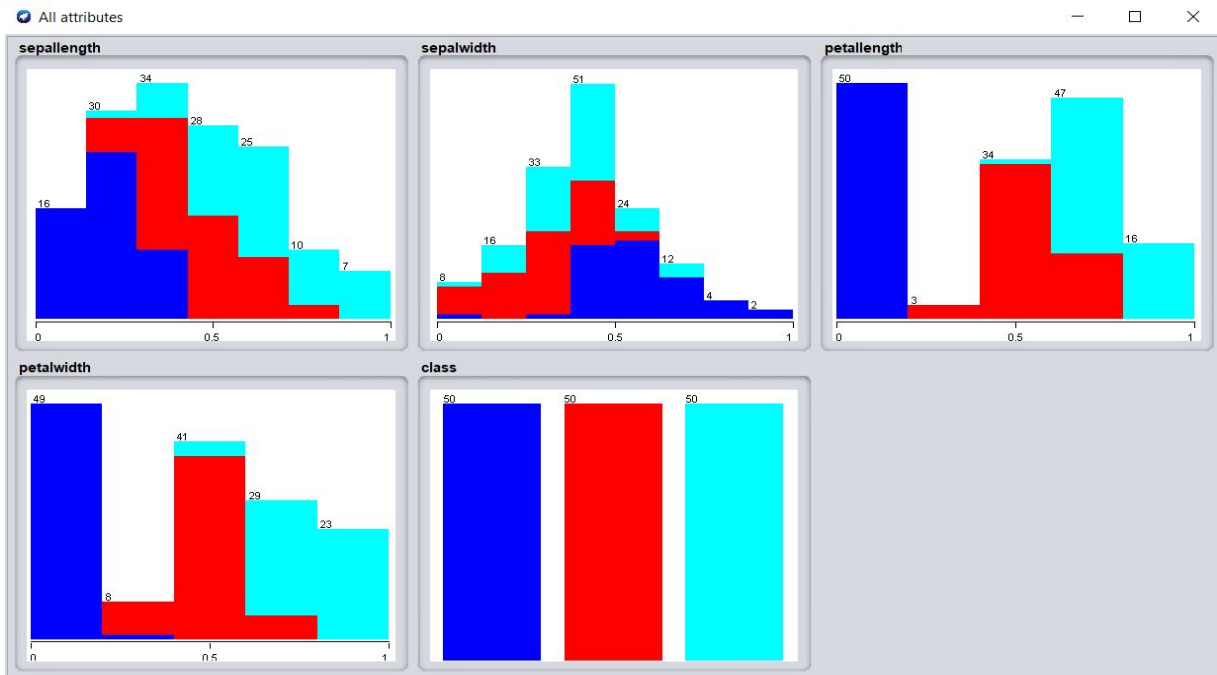
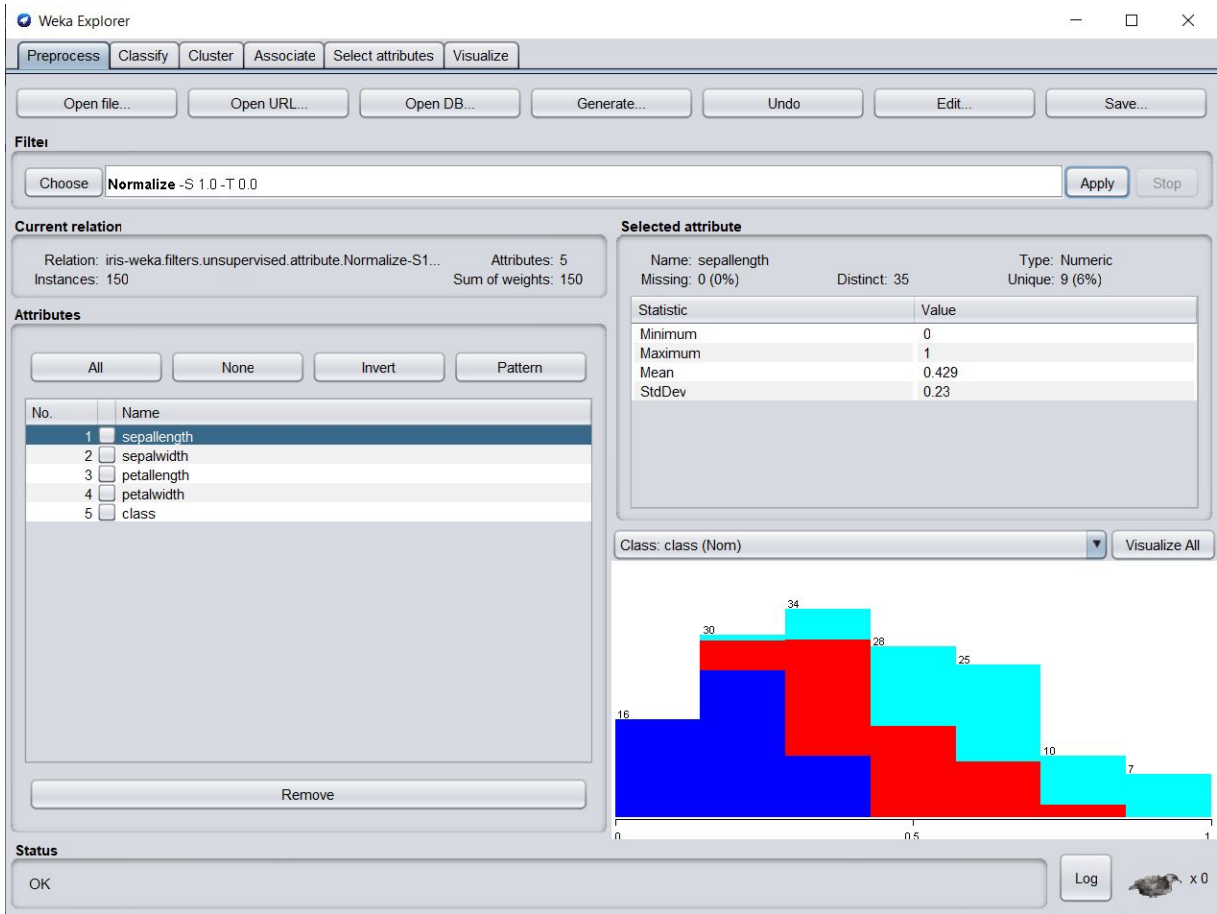
Agrupamento de Dados

01 - Iris

Base de datos Iris antes de aplicar os Filtros.

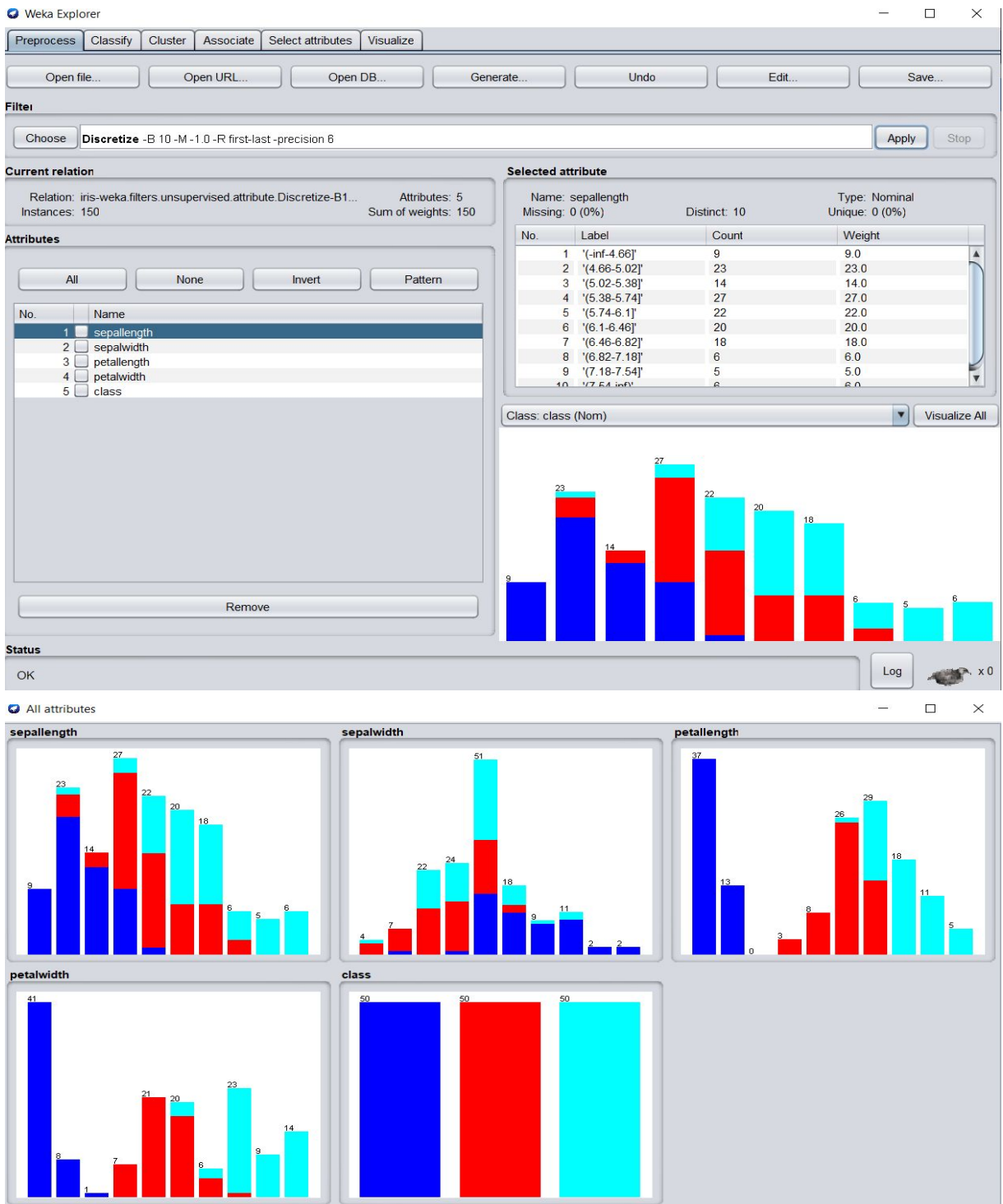


Após aplicar Normalização.



- Podemos observar que tivemos os dados mais bem distribuídos quando nos atentamos para os atributos, máximo, mínimo e média. Graficamente não se alterou.

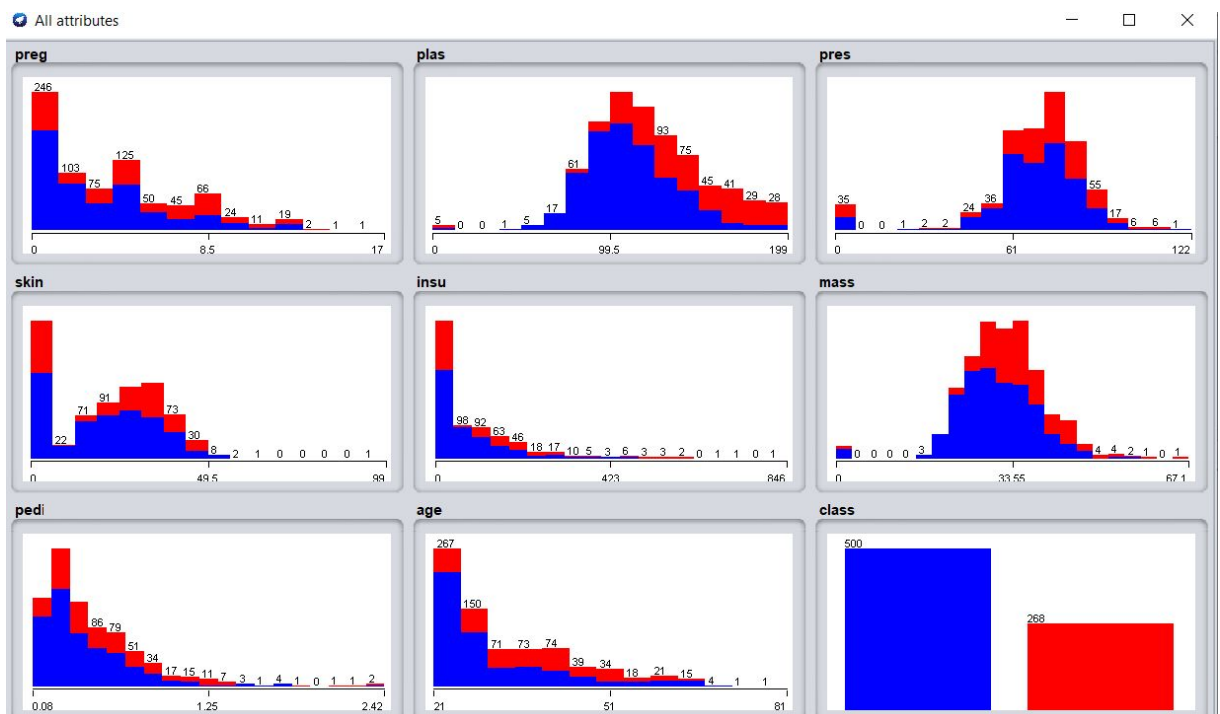
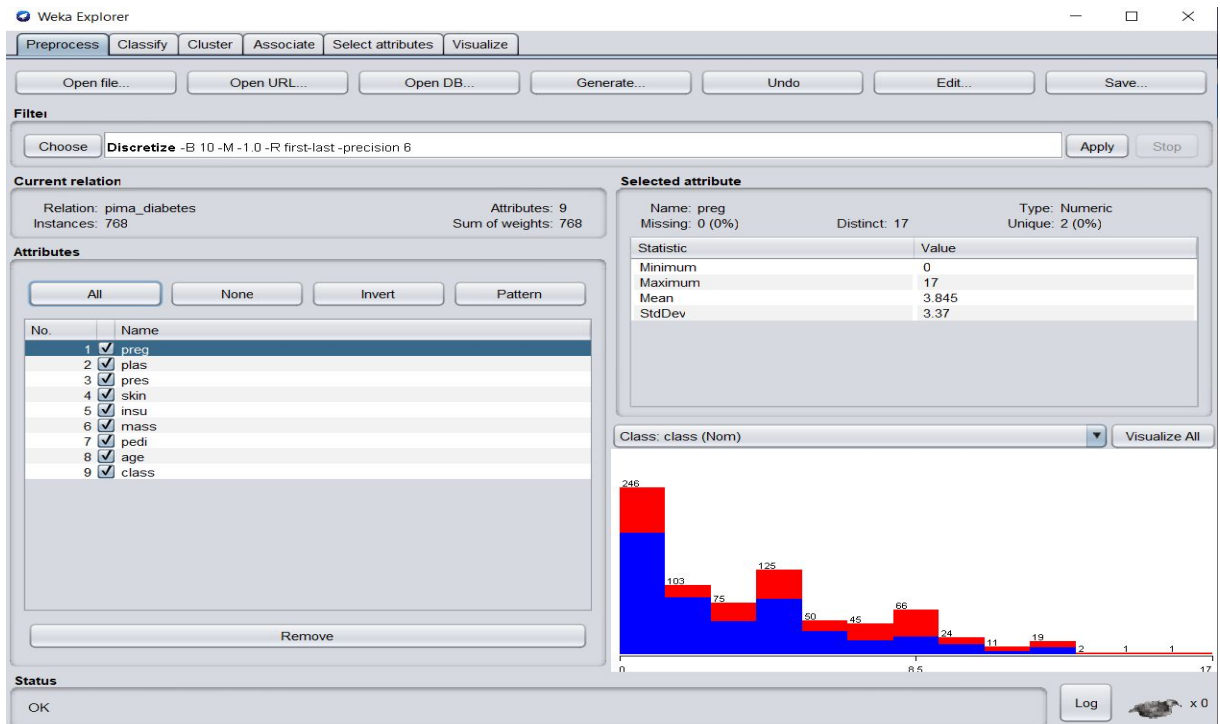
Após aplicar Discretização:



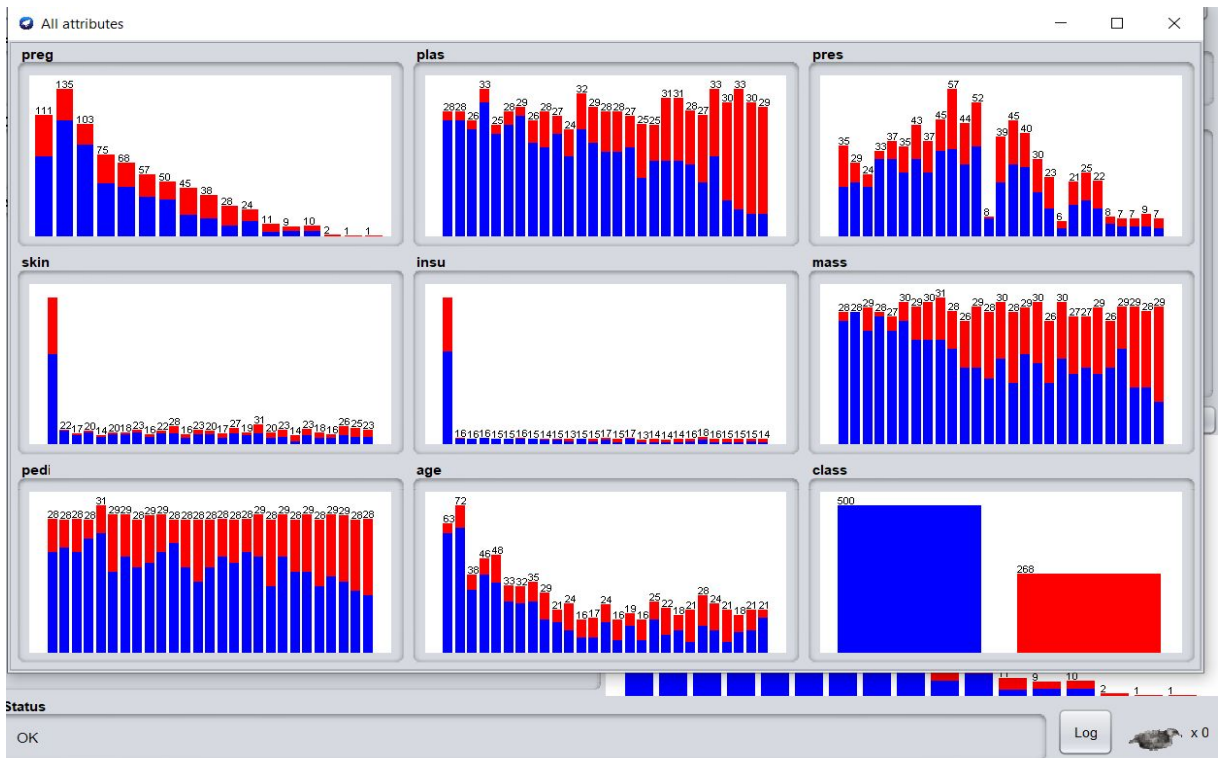
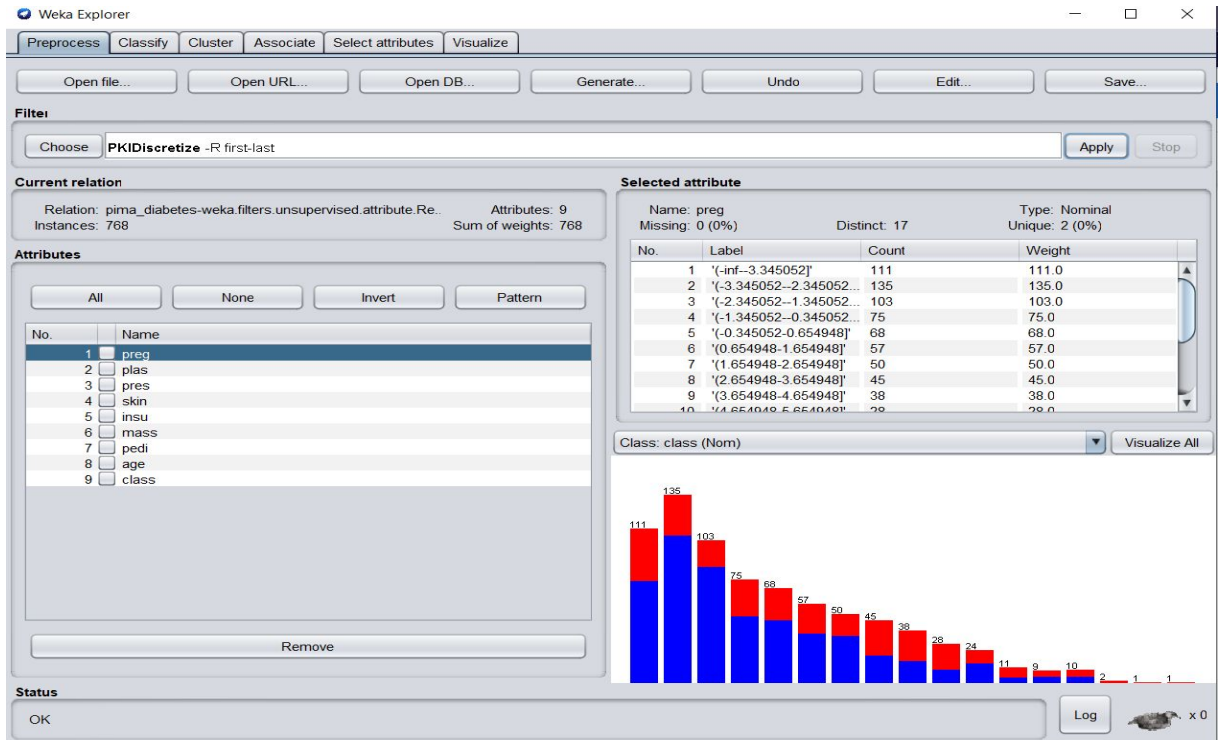
- Podemos observar que os dados foram discretizados, tendo uma maior limitação dos valores, graficamente separados em stacked bars. Quando olhamos para a coluna sepallength podemos observar claramente como os dados foram discretizados

02 - Diabetes

Base de dados antes de aplicar os filtros



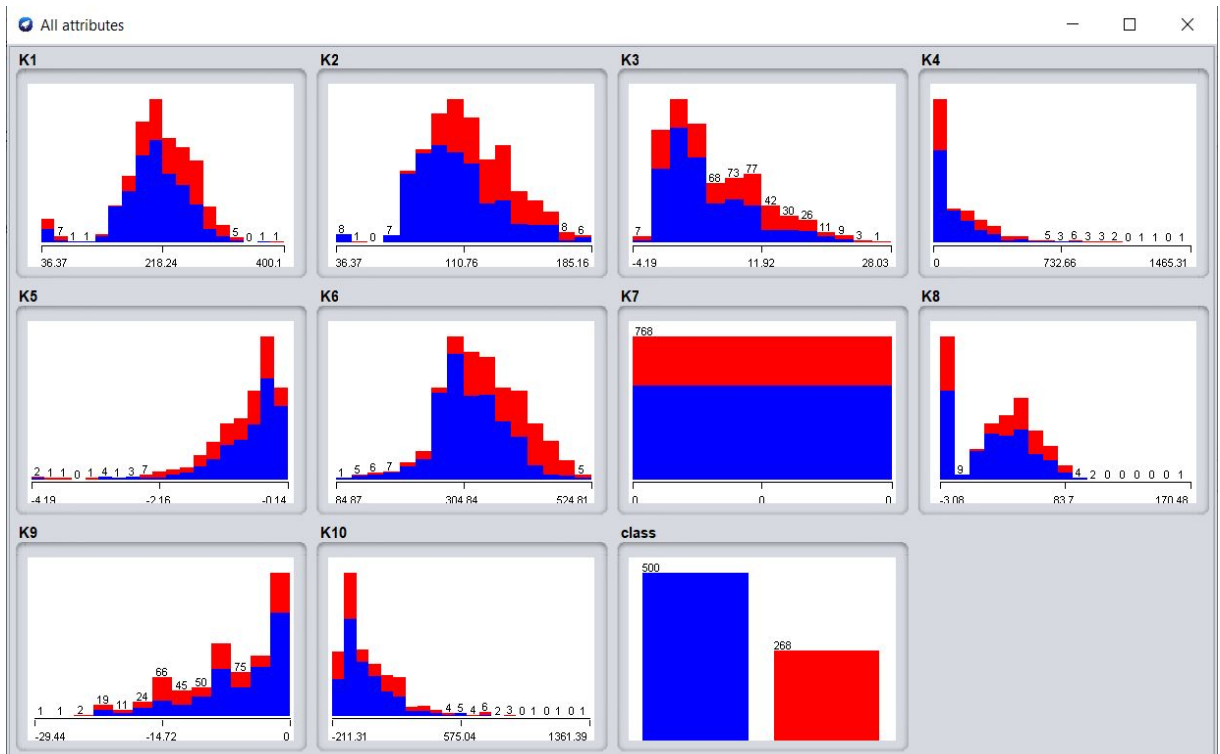
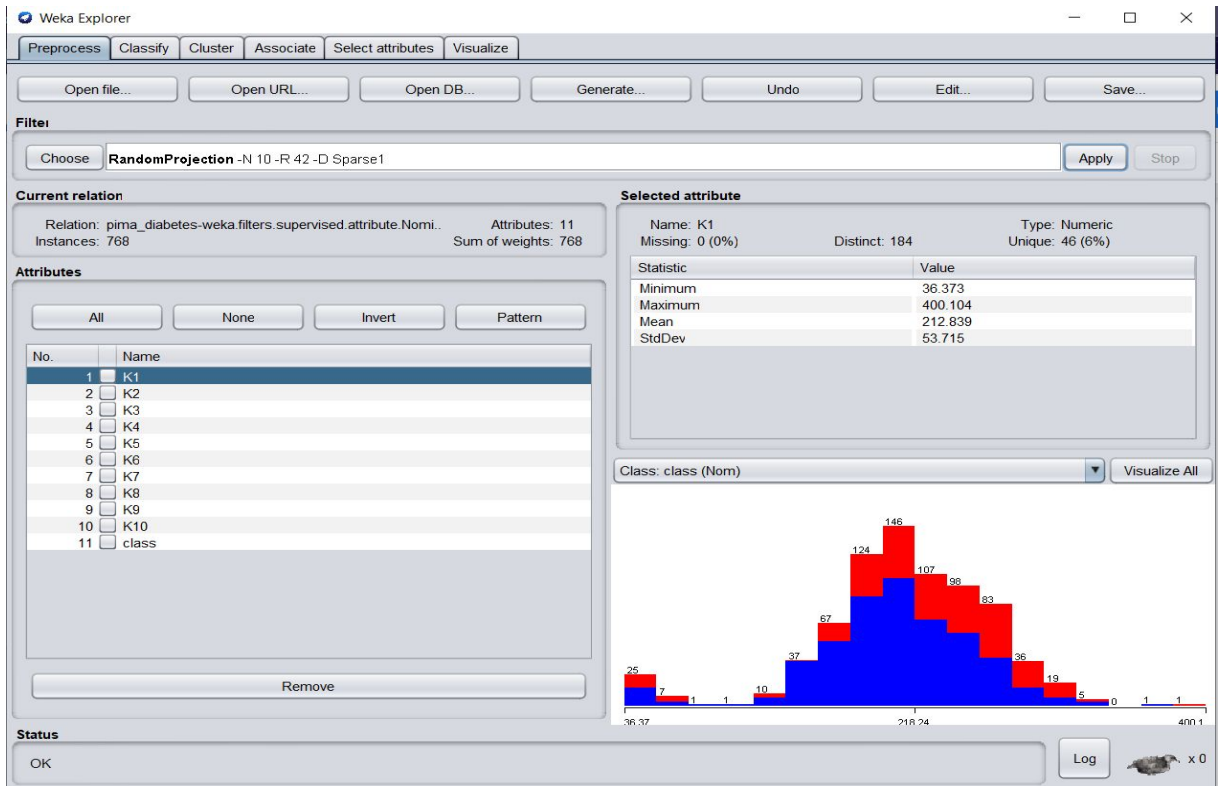
Após aplicar PKIDiscretize:



- Podemos observar que os dados foram discretizados, tendo uma maior limitação dos valores, graficamente separados em stacked bars. Quando olhamos para a coluna preg podemos observar claramente como os dados

foram discretizados, usando outra algoritmo de discretização porém chegando a um resultado semelhante ao usado na base de dados Iris

Após aplicar RandomProjection:



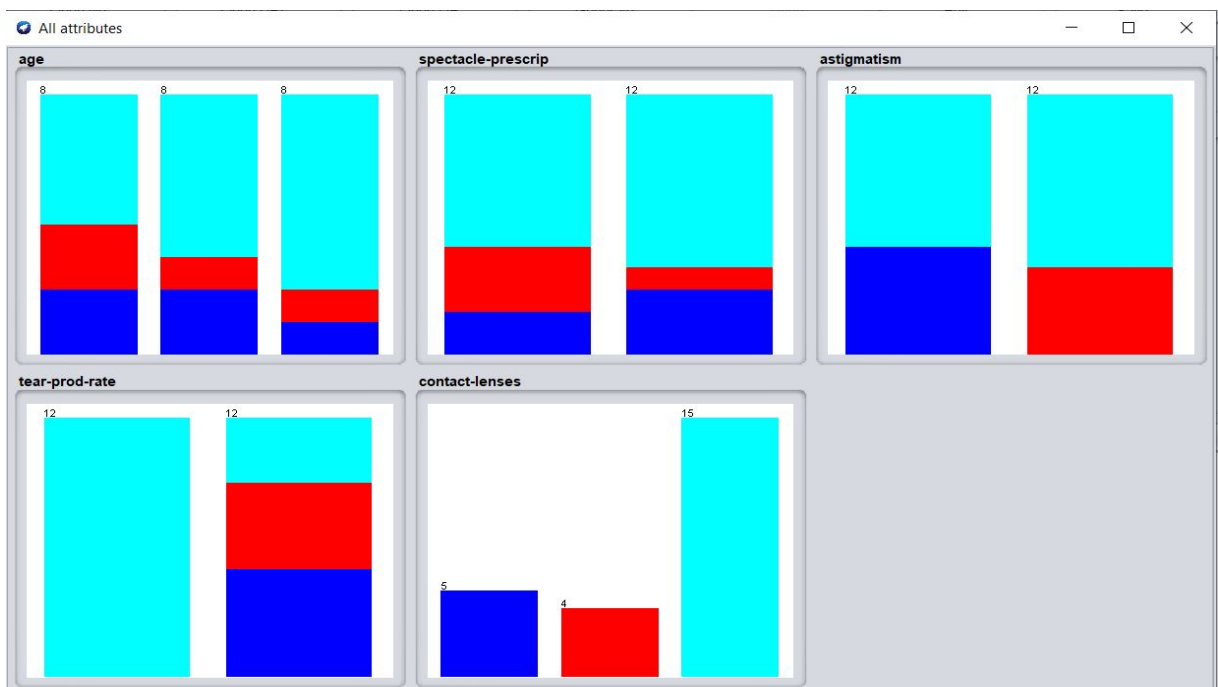
- Pode-se observar que houve uma redução na dimensionalidade dos dados, ou seja diminui-se também os valores contidos no dataset, deixando-o mais limpo. Fica bem visível essa mudança graficamente.

03 - Contact Lenses

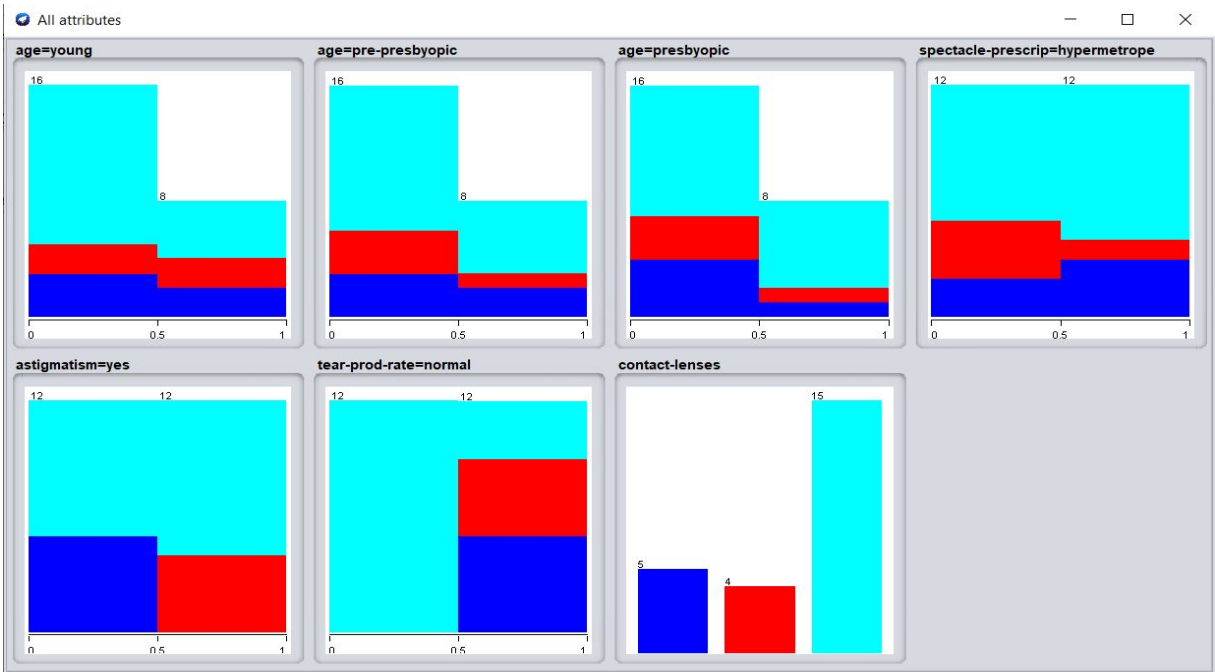
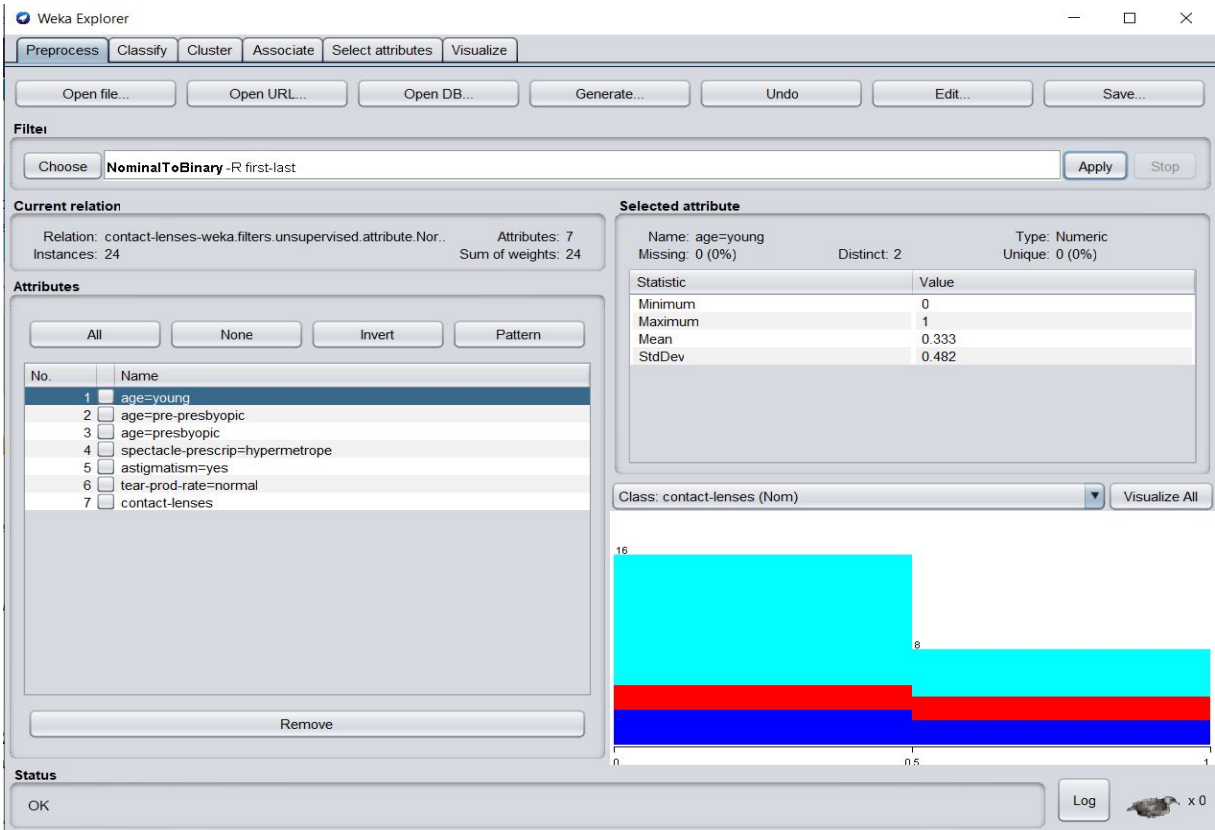
Antes de aplicar os Filtros:

Weka Explorer interface showing the 'Visualize' tab. The 'Attributes' list on the left includes 'age', 'spectacle-prescrip', 'astigmatism', 'tear-prod-rate', and 'contact-lenses'. The 'Selected attribute' panel on the right shows details for 'age' (Nominal, 3 distinct values) and a table of counts and weights for 'young', 'pre-presbyopic', and 'presbyopic'. Below this, three small stacked bar charts are visible, representing the distribution of the 'age' attribute across the three classes of 'contact-lenses'.

No.	Label	Count	Weight
1	young	8	8.0
2	pre-presbyopic	8	8.0
3	presbyopic	8	8.0

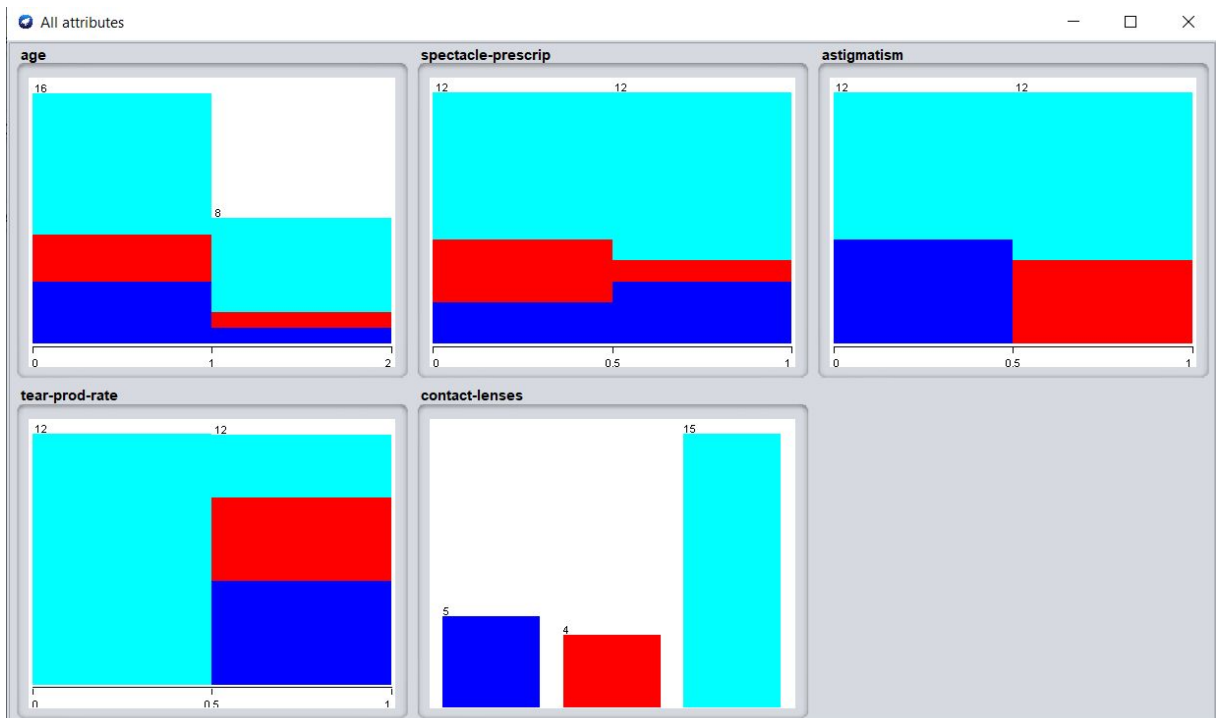
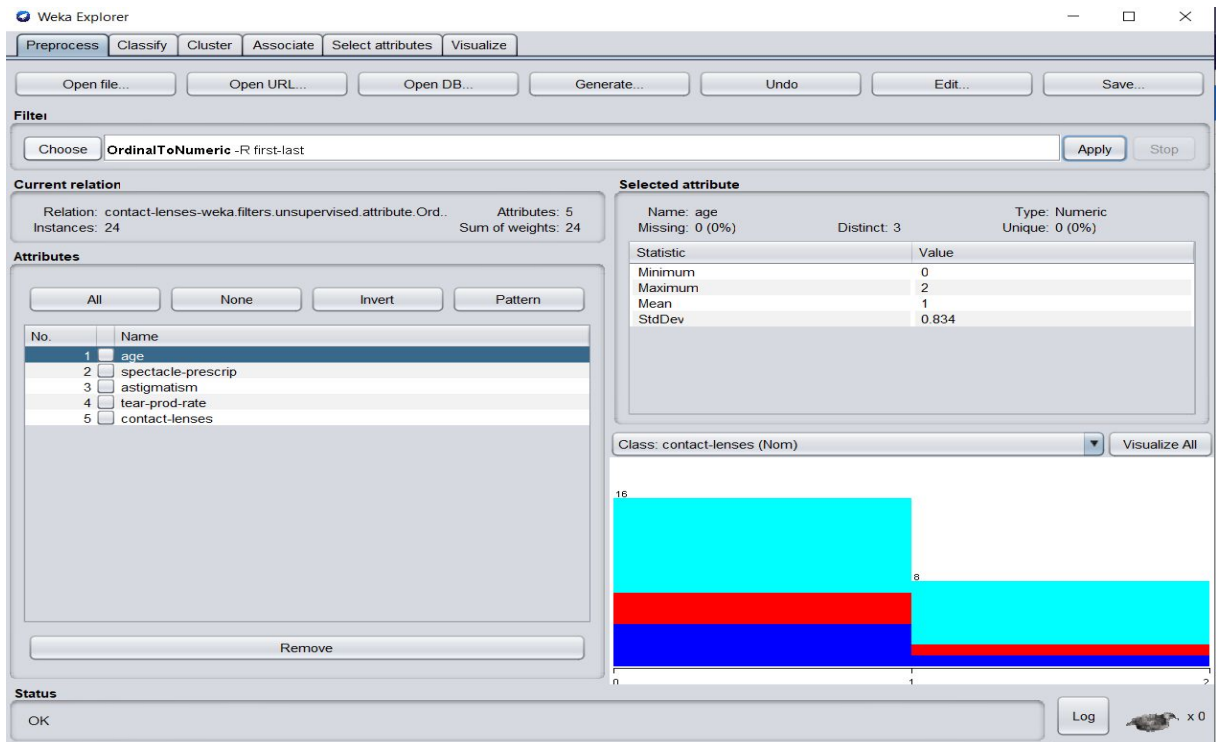


Após aplicar NominalToBinary



- Podemos observar que houve uma binarização dos dados, trocando até mesmo os nomes das colunas, graficamente observa se que houve uma variável que se estendeu mais dos que as outras justamente por se encaixar melhor no atributo nominal.

Após aplicar OrdinalToNumeric:



- Podemos observar que houve uma troca de valores na base de dados e houve uma variável que se expandiu novamente por se encaixar melhor ao filtro.

1 - d)

Kmeans - Iris - 3 clusters

=== Run information ===

```
Scheme:          weka.clusterers.SimpleKMeans -init 0 -max-candidates 100
-periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:  iris-weka.filters.unsupervised.attribute.Remove-R5
Instances:  150
Attributes:  4
              sepallength
              sepalwidth
              petallength
              petalwidth
Test mode:   evaluate on training data
```

=== Clustering model (full training set) ===

kMeans

=====

Number of iterations: 6

Within cluster sum of squared errors: 6.998114004826762

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4

Cluster 1: 6.2,2.9,4.3,1.3

Cluster 2: 6.9,3.1,5.1,2.3

Missing values globally replaced with mean/mode

Final cluster centroids:

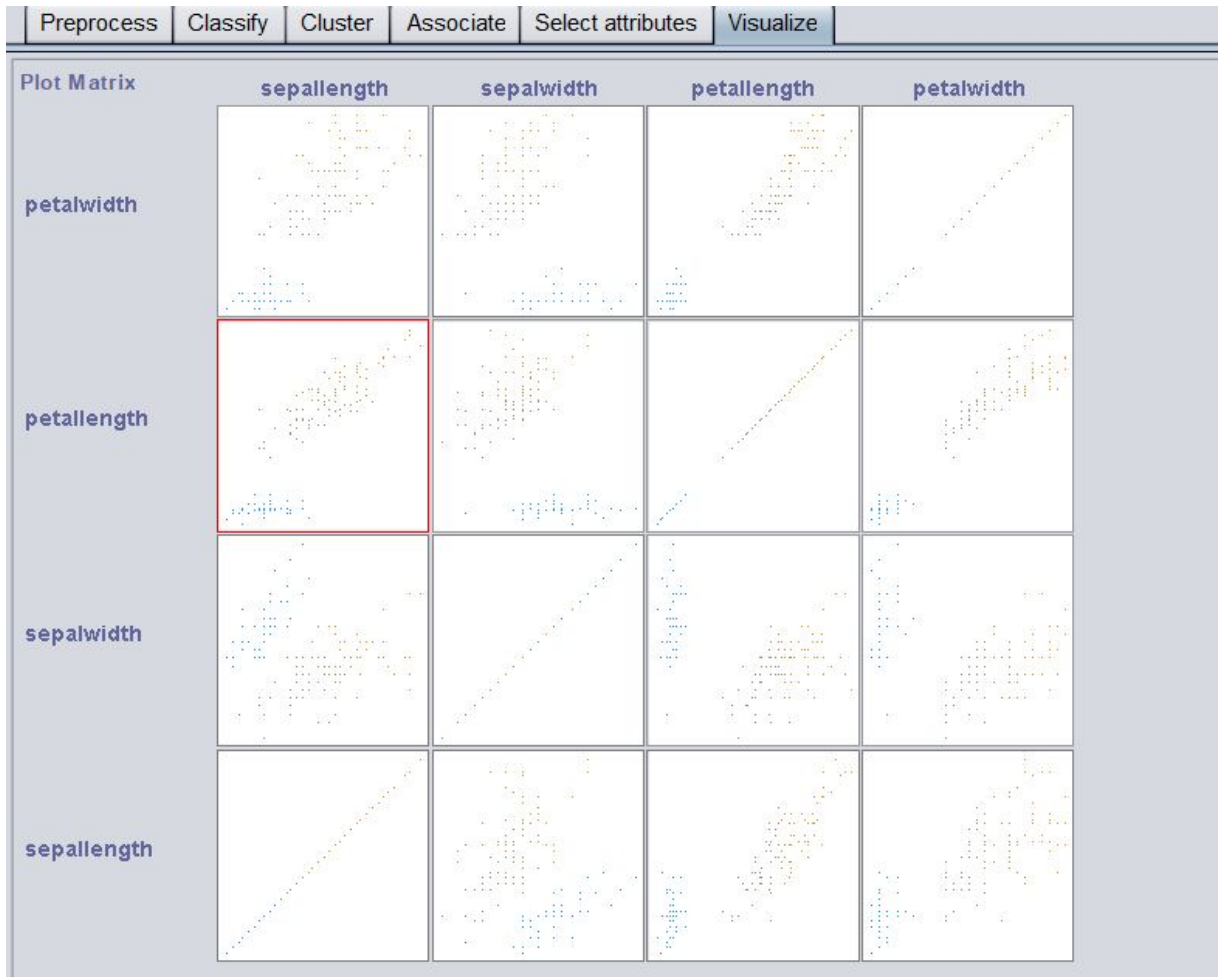
Attribute	Cluster#			
	Full Data	0	1	2
	(150.0)	(61.0)	(50.0)	(39.0)
=====				
sepal length	5.8433	5.8885	5.006	6.8462
sepal width	3.054	2.7377	3.418	3.0821
petal length	3.7587	4.3967	1.464	5.7026
petal width	1.1987	1.418	0.244	2.0795

Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	61 (41%)
1	50 (33%)
2	39 (26%)



Kmeans - Diabetes - 2 Clusters

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100
 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A
 "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
 Relation: pima_diabetes-weka.filters.unsupervised.attribute.Remove-R9
 Instances: 768
 Attributes: 8
 preg
 plas
 pres
 skin

insu
mass
pedi
age

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

=====

Number of iterations: 7

Within cluster sum of squared errors: 121.2579017999101

Initial starting points (random):

Cluster 0: 1,126,56,29,152,28.7,0.801,21

Cluster 1: 8,95,72,0,0,36.8,0.485,57

Missing values globally replaced with mean/mode

Final cluster centroids:

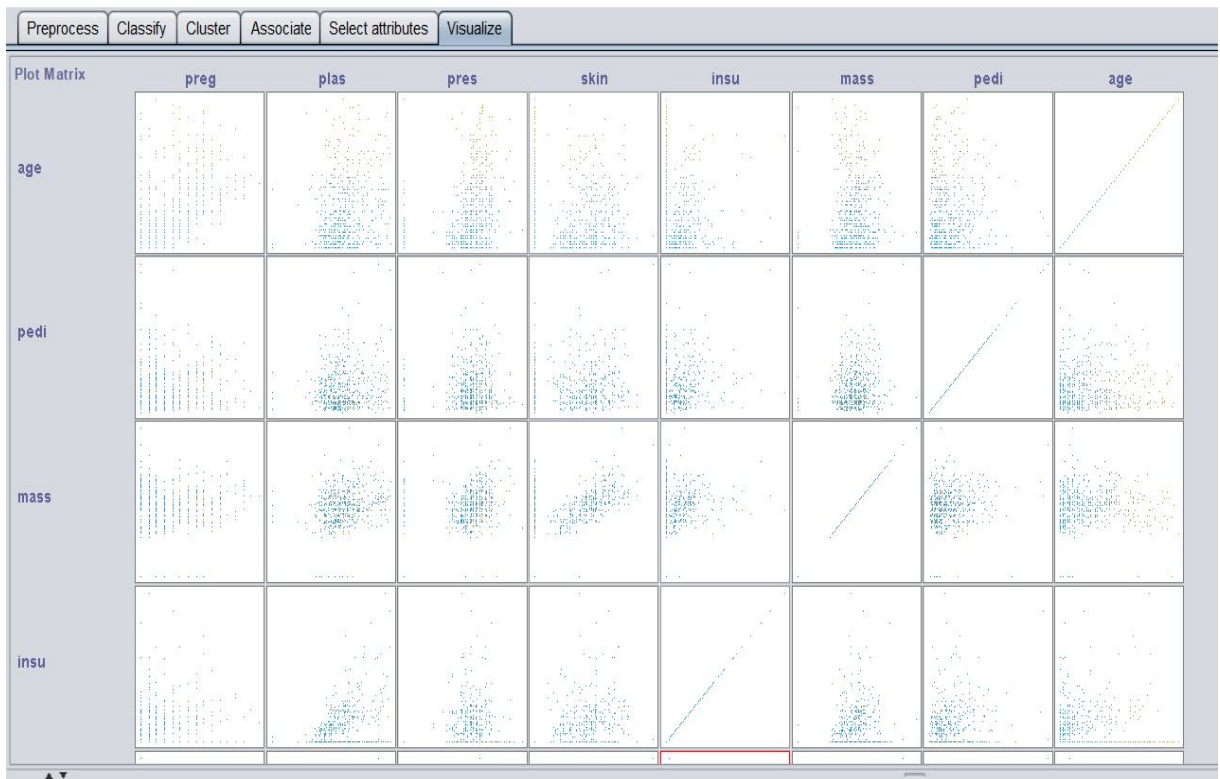
Attribute	Cluster#		
	Full Data	0	1
	(768.0)	(515.0)	(253.0)
=====			
preg	3.8451	2.0835	7.4308
plas	120.8945	115.3282	132.2253
pres	69.1055	65.9903	75.4466
skin	20.5365	21.8194	17.9249
insu	79.7995	85.0194	69.1739
mass	31.9926	31.7751	32.4352
pedi	0.4719	0.4708	0.4741
age	33.2409	26.7728	46.4071

Time taken to build model (full training data) : 0.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 515 (67%)
1 253 (33%)



Kmeans - Glass - 4 Clusters

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100
-periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 4 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation: Glass-weka.filters.unsupervised.attribute.Remove-R10

Instances: 214

Attributes: 9

RI

Na

Mg

Al

Si

K

Ca

Ba

Fe

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

=====

Number of iterations: 8

Within cluster sum of squared errors: 26.28430433007474

Initial starting points (random):

Cluster 0: 1.52152,13.05,3.65,0.87,72.32,0.19,9.85,0,0.17

Cluster 1: 1.51618,13.53,3.55,1.54,72.99,0.39,7.78,0,0

Cluster 2: 1.51316,13.02,0,3.04,70.48,6.21,6.96,0,0

Cluster 3: 1.51625,13.36,3.58,1.49,72.72,0.45,8.21,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:

	Cluster#				
Attribute	Full Data	0	1	2	3
	(214.0)	(40.0)	(46.0)	(2.0)	(126.0)
=====					
==					
RI	1.5184	1.5182	1.5194	1.5132	1.5181
Na	13.4079	13.0907	13.937	13.01	13.3217
Mg	2.6845	3.4073	0.1539	0	3.4216
Al	1.4449	1.2765	1.7778	3.03	1.3517

Si	72.6509	72.6908	72.9291	70.59	72.5694
K	0.4971	0.512	0.218	6.21	0.5035
Ca	8.957	8.7575	10.1867	6.945	8.6033
Ba	0.175	0.0253	0.6628	0	0.0473
Fe	0.057	0.219	0.0411	0	0.0123

Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	40 (19%)
1	46 (21%)
2	2 (1%)
3	126 (59%)

