Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Computação

Alan L. Melo - 12323CCP001
Alessandro S. Angeruzzi - 12322CCP001
Paulo Victor da S. Freitas - 12323CCP011

Aula Prática 2:
Ferramentas Weka

Disciplina: Agrupamento de Dados
Professores.:
- Dra. Elaine Ribeiro Faria
- Dr. Bruno Augusto Nassif Travençolo
- Dr. Rafael Dias Araújo

Uberlândia
2024

1. Bases de Dados

Para a execução da atividade proposta foram selecionadas 3 bases de dados relacionadas a seguir:

Rice (Cammeo and Osmancik) (https://archive.ics.uci.edu/dataset/545/rice+cammeo+and+osmancik)
Base com 3.810 registros contendo, 7 características morfológicas de grãos de arroz de 2 espécies, Cammeo e Osmancik.

Wine Quality, red (https://archive.ics.uci.edu/dataset/186/wine+quality)
Base com 4898 registros, com 11 características de vinhos e um score de qualidade.

Heart Disease, Cleveland (https://archive.ics.uci.edu/dataset/45/heart+disease)
Base com 303 registros, contendo 13 características de pacientes e indicação se possui doença cardíaca.

2. Aplicação de Filtros

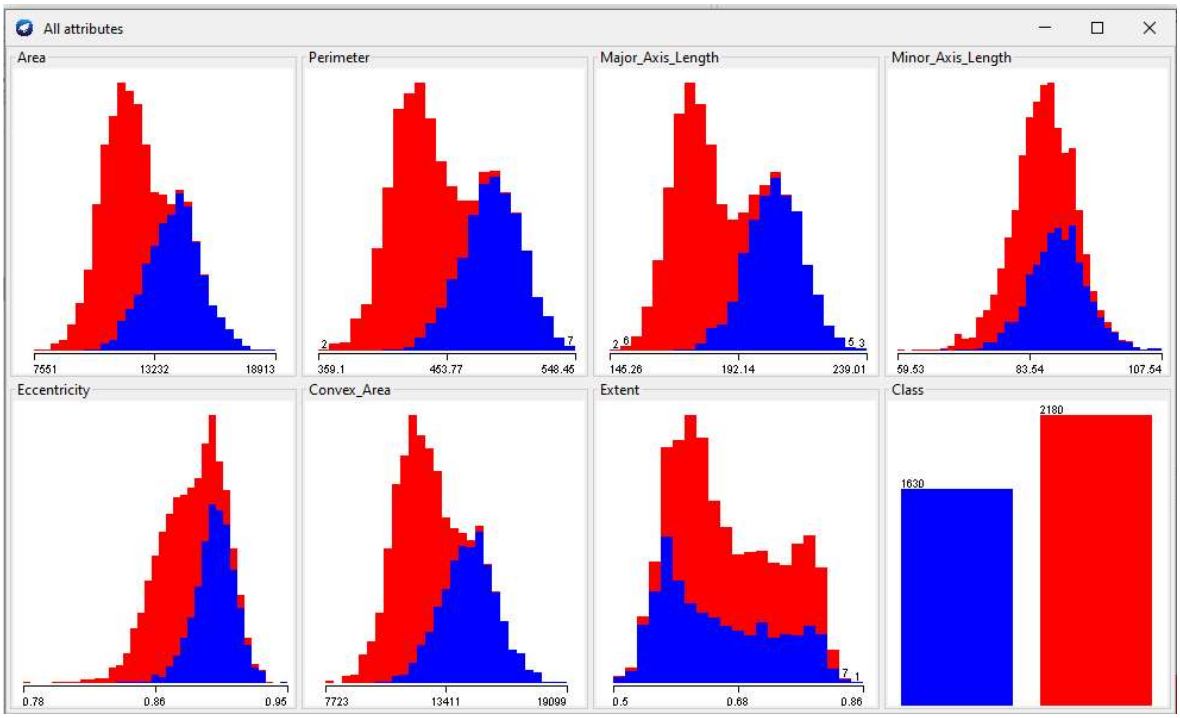
Em cada uma das 3 bases foram aplicados 2 filtros e evidenciadas as alterações que ocorreram nos atributos em comparação com os valores originais.

2.1. Base Rice

Aplicação dos filtros na base Rice (Cammeo and Osmancik).

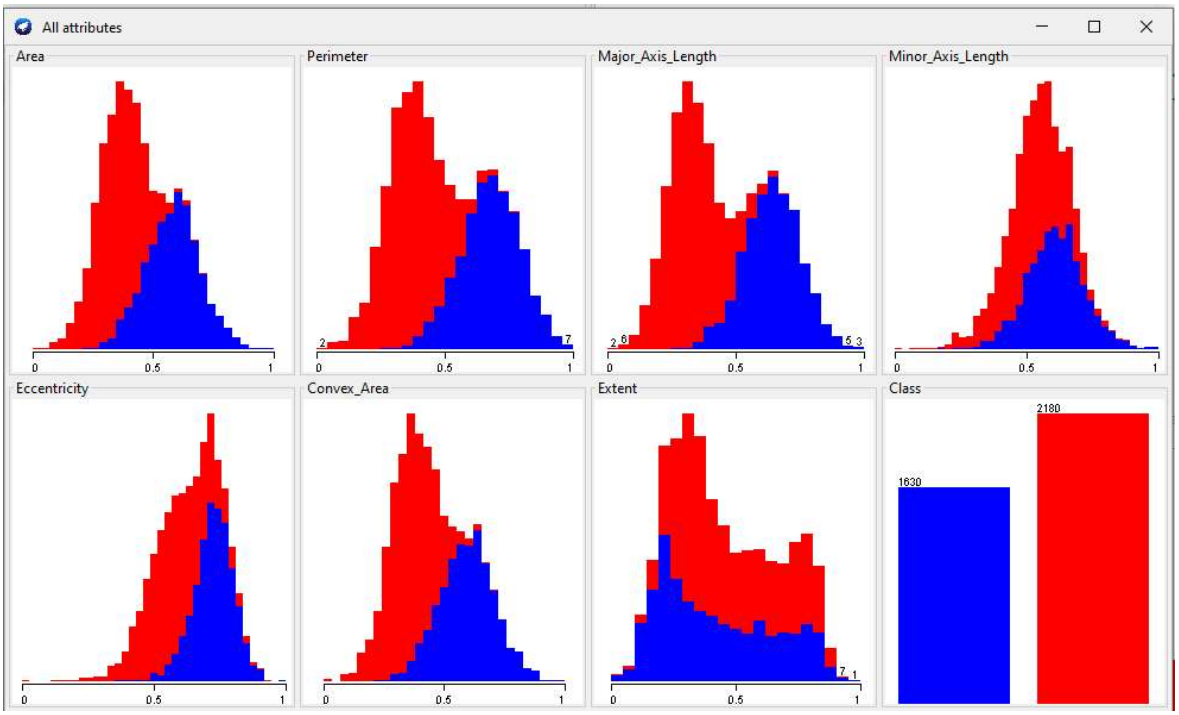
2.1.1. Atributos Originais

Registro das distribuições dos atributos antes da aplicação dos filtros. Como a base possui um atributo de classificação binário (Class), indicando a espécie do arroz, também podemos ver nas distribuições essa classificação.



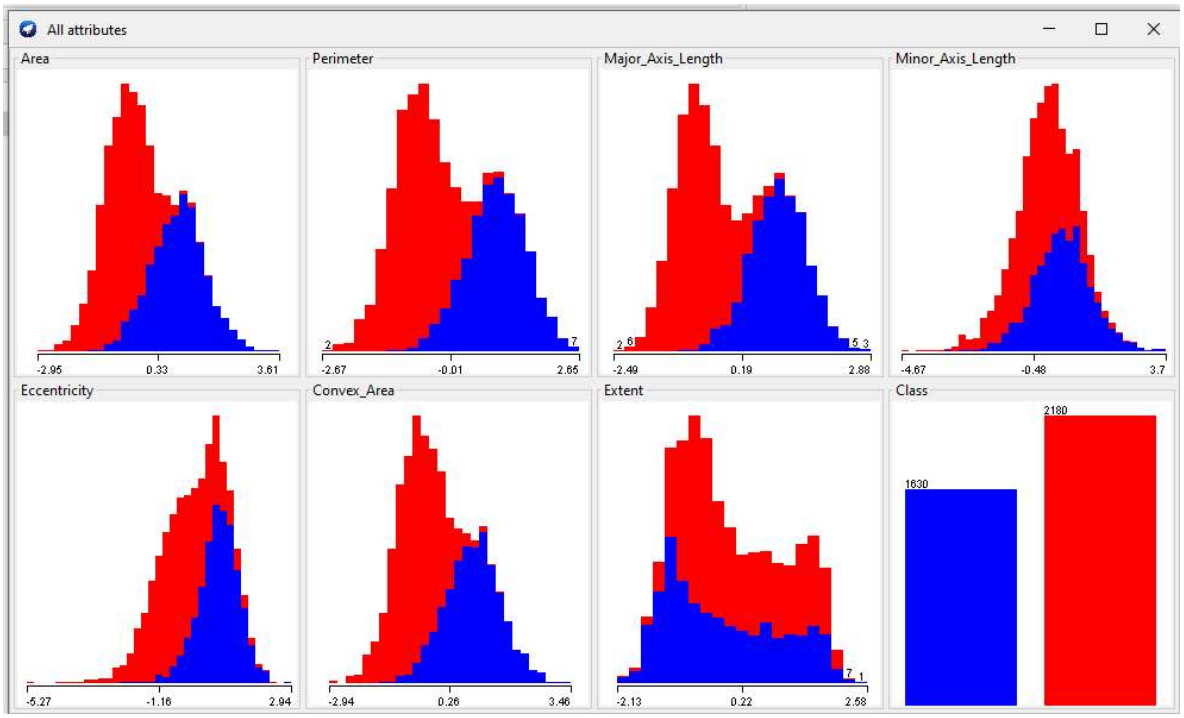
2.1.2. Filtro Normalize

Como os dados possuem escalas diferentes, passa a ser interessante a aplicação do filtro Normalize, ou normalização em português, onde os valores são convertidos para o intervalo de 0 a 1, tendo assim todos os atributos dentro de uma mesma escala comparável. Nesta transformação podem ocorrer mudanças na forma de distribuição dos dados, porém não foi observado neste caso.



2.1.3. Filtro Standardize

O filtro Standardize, ou padronização em português, assim como o filtro anterior é aplicado para transformar os valores dos atributos trazendo para uma mesma escala e com a vantagem de não alterar a forma de distribuição dos dados; O filtro transforma os dados para que tenham média zero e desvio padrão igual a 1.

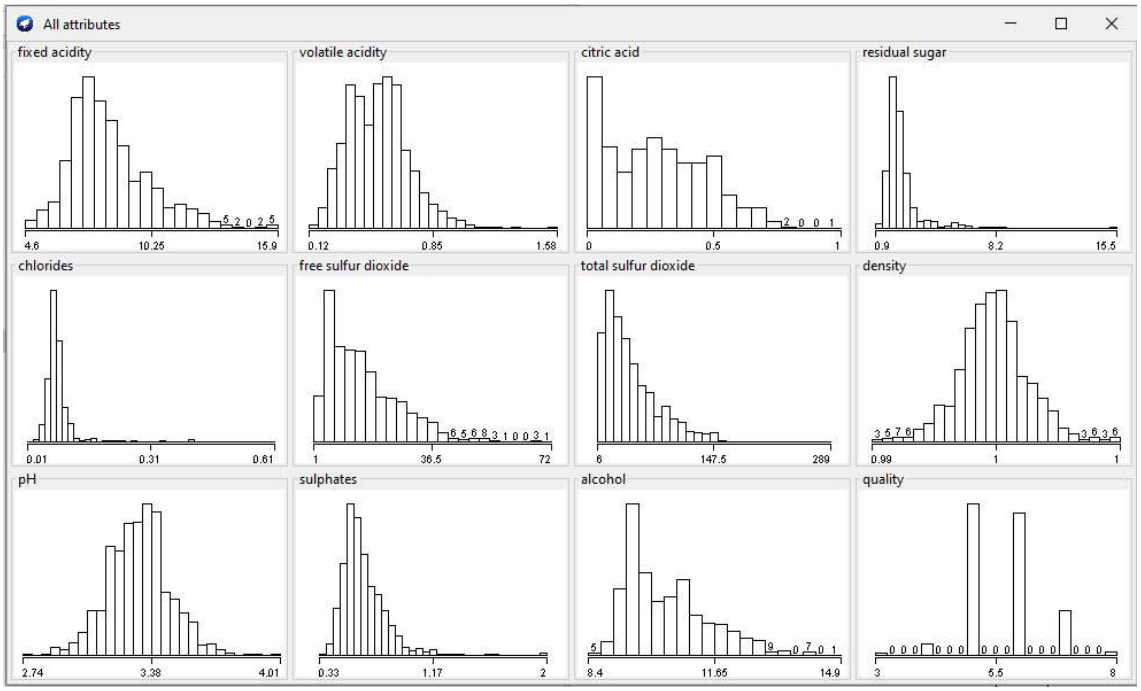


2.2. Wine Quality

Aplicação dos filtros na Wine Quality, red.

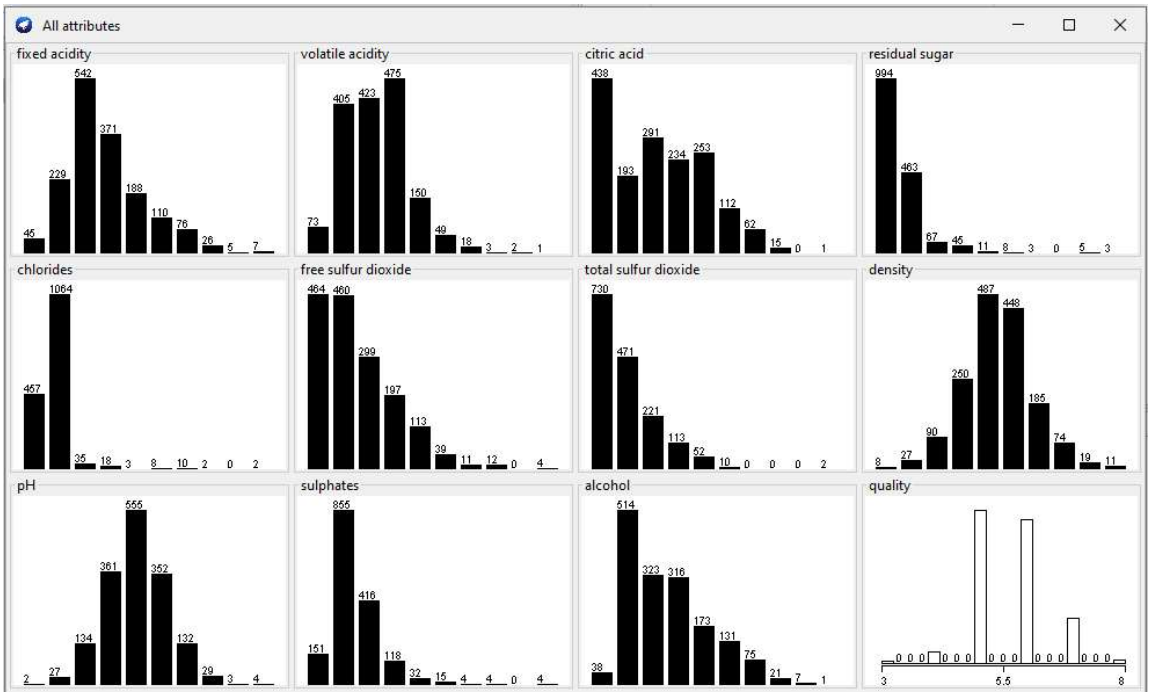
2.2.1. Atributos Originais

Registro das distribuições dos atributos antes da aplicação dos filtros. O atributo “quality” é um label de classificação que varia de 0 a 10.



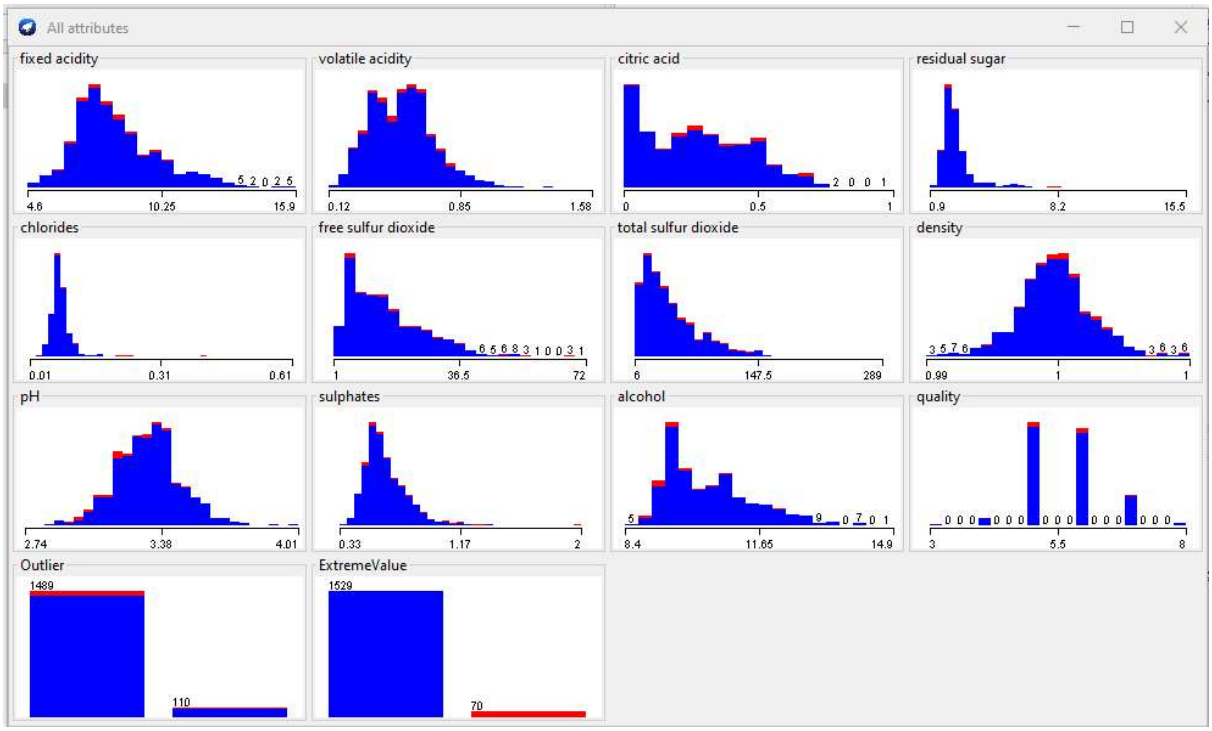
2.3. Filtro Discretize

Sendo os dados da base medições com valores contínuos, o filtro Discretize, ou Discretização, passa a ser interessante para agrupar os valores em intervalos, apesar de não ser uma transformação recomendada para aplicação do K-means.



2.4. Filtro Interquartile Range

Ao observar as distribuições da base levantasse a suspeita da existência de Outliers que podem afetar de forma significativa o resultado do K-means por exemplo, por isso a aplicação do Interquartile Range é interessante para a identificação de Outliers e Extreme Values para se avaliar uma ação de tratamento destes.

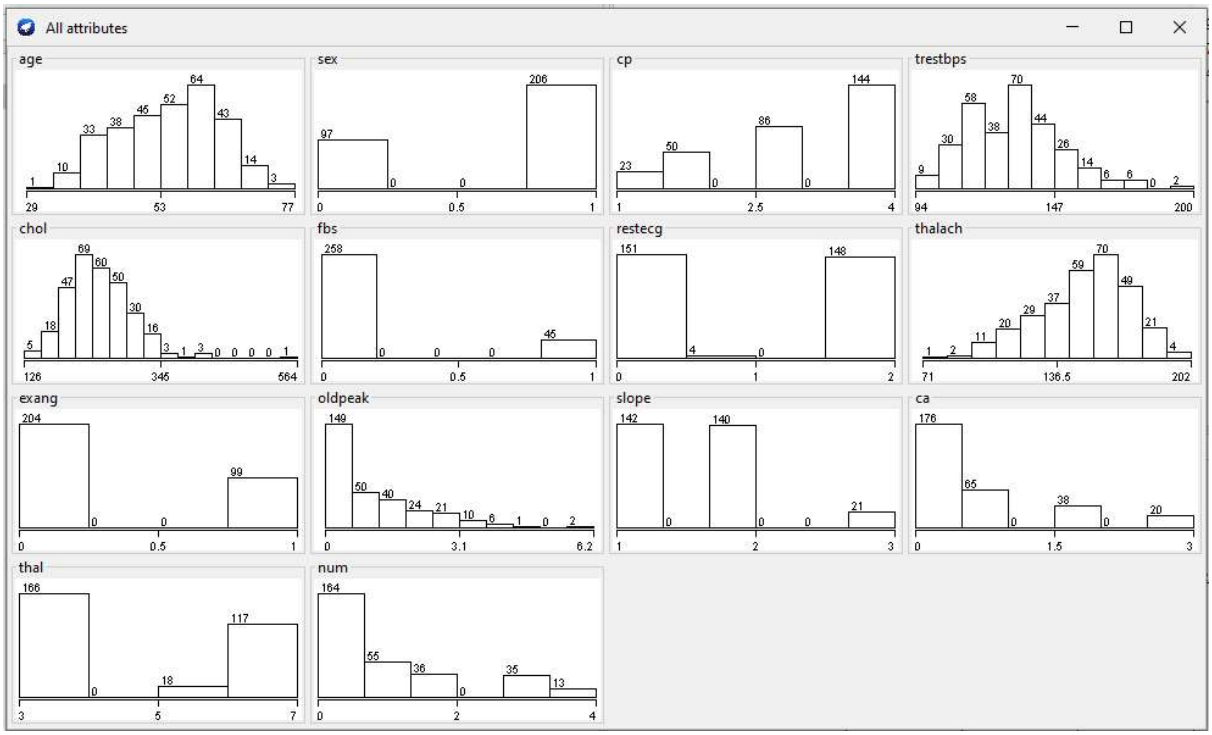


3. Heart Disease

Aplicação dos filtros na base Heart Disease, Cleveland.

3.1. Atributos Originais

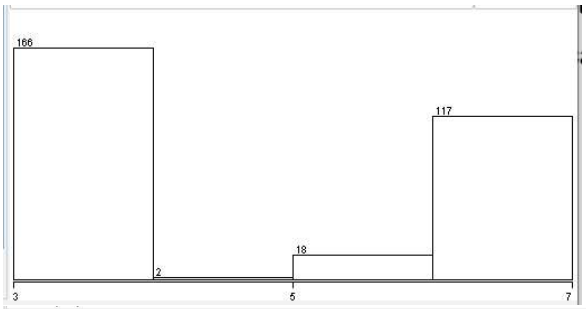
Registro das distribuições dos atributos antes da aplicação dos filtros. O atributo “num” é um label de classificação que indica o diagnóstico de doença cardíaca e varia de 0 a 4, sendo 0 a ausência e valores maiores que 1 um diagnóstico positivo.



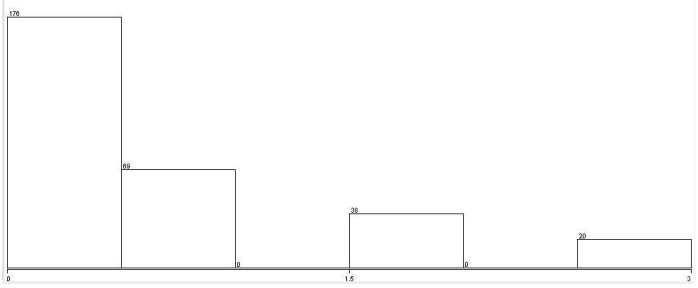
3.2. Filtro Missing Values

Apesar de poucos, a base possui alguns registros com dados faltantes, dando a oportunidade de uso desse filtro. Havia 2 registros sem valores no atributo “thal” e 4 registros no atributo “ca”:

Nova Distribuição de “thal”:



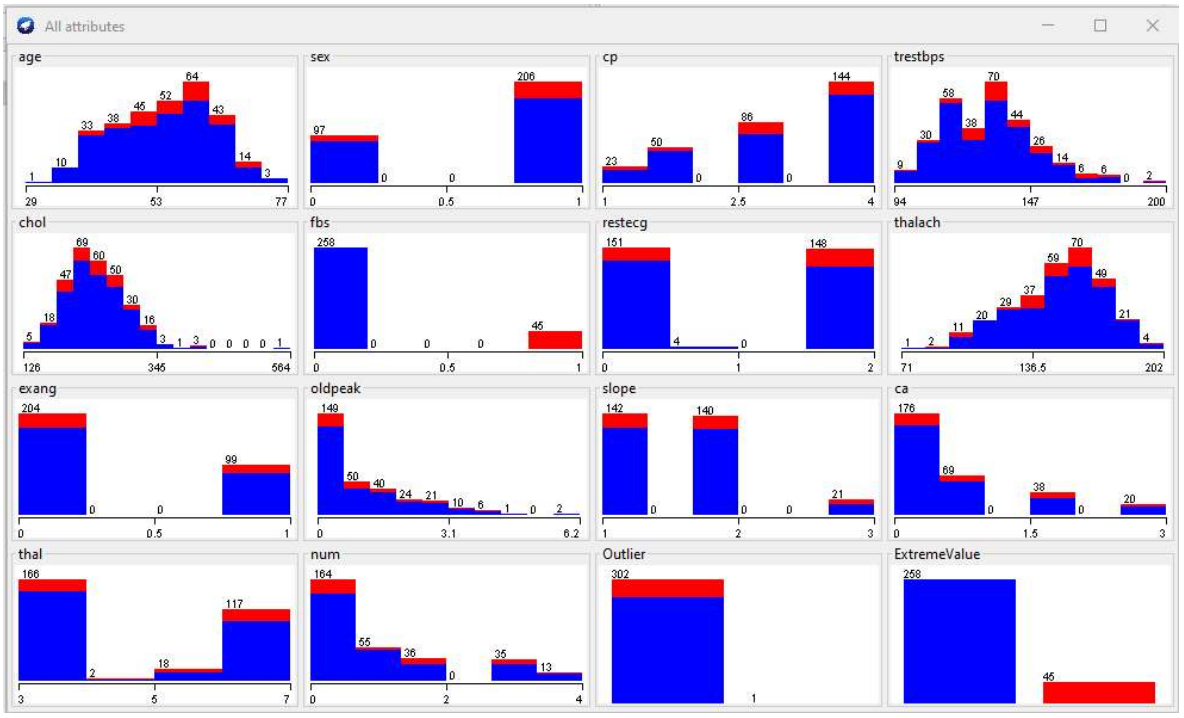
Nova Distribuição de “ca”:



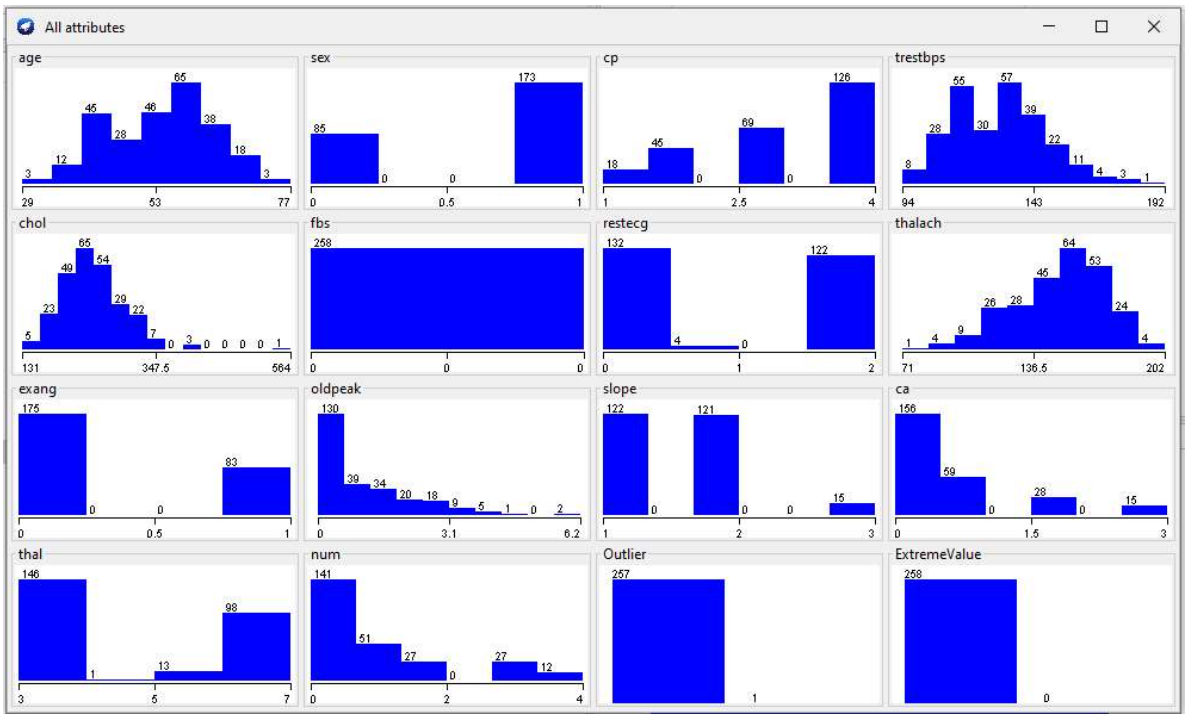
3.3. Filtro Remove With Values

Como a base possui Outliers podemos aplicar algum tratamento para estes dados, uma opção é a utilização do filtro Remove With Values após a aplicação do filtro Interquartile Range que identifica os Outliers e Valores Extremos como foi feito na base Wine Quality.

Primeiramente aplica-se o filtro Interquartile Range:



E depois o filtro Remove With Values para os atributos Outlier e Extreme Value que foram gerados:



4. Aplicação do K-means

O algoritmo K-means será aplicado nas 3 bases com diferentes valores de seed.

4.1. Base Rice

Aplicação do algoritmo K-means na base Rice (Cammeo and Osmancik) com a geração de 2 clusters.

As execuções com diferentes Seeds foram finalizadas com centróides diferentes e com isso também uma diferença entre os clusters. Porém a diferença foi pequena e não relevante na validação dos clusters utilizando a classe pré existente.

Execução com Seed = 10

```
Number of iterations: 8
Within cluster sum of squared errors: 381.62723306622354

Initial starting points (random):

Cluster 0: 11682,437.040009,176.230988,86.322495,0.87182,11969,0.581194
Cluster 1: 15172,504.15799,213.22467,91.667061,0.902873,15477,0.622441

Missing values globally replaced with mean/mode

Final cluster centroids:
                            Cluster#
Attribute            Full Data          0          1
                      (3810.0)   (2232.0)   (1578.0)
=========================================================
Area                12667.7276 11478.1389 14350.3397
Perimeter             454.2392   428.5994   490.5054
Major_Axis_Length     188.7762   176.2258   206.5281
Minor_Axis_Length      86.3138    84.0274    89.5477
Eccentricity            0.8869     0.8774     0.9002
Convex_Area         12952.4969 11727.5484 14685.1236
Extent                  0.6619     0.6712     0.6488

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

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

Clustered Instances

0      2232 ( 59%)
1      1578 ( 41%)


Class attribute: Class
Classes to Clusters:

    0    1  <-- assigned to cluster
  188 1442 | Cammeo
 2044  136 | Osmancik

Cluster 0 <-- Osmancik
Cluster 1 <-- Cammeo

Incorrectly clustered instances :     324.0     8.5039 %
```

Execução com Seed = 42

```
Number of iterations: 13
Within cluster sum of squared errors: 381.62701696675845

Initial starting points (random):

Cluster 0: 12653,441.493988,176.015106,93.187904,0.848353,12955,0.764809
Cluster 1: 11999,440.669006,176.701904,88.940865,0.864089,12483,0.646846

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute            Full Data          0          1
                      (3810.0)   (1580.0)   (2230.0)
=====================================================
Area                12667.7276 14348.2297 11477.0578
Perimeter             454.2392   490.4623   428.5743
Major_Axis_Length     188.7762   206.5109   176.2109
Minor_Axis_Length      86.3138    89.5416    84.0268
Eccentricity            0.8869     0.9002     0.8774
Convex_Area         12952.4969 14682.8956 11726.4744
Extent                  0.6619     0.6488     0.6713


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

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

Clustered Instances

0      1580 ( 41%)
1      2230 ( 59%)


Class attribute: Class
Classes to Clusters:

    0    1  <-- assigned to cluster
 1443  187 | Cammeo
  137 2043 | Osmancik

Cluster 0 <-- Cammeo
Cluster 1 <-- Osmancik

Incorrectly clustered instances :     324.0     8.5039 %
```

4.2. Wine Quality

Aplicação do algoritmo K-means na base Wine Quality, red, com a geração de 6 clusters.

Mesmo com a mudança do seed e a inicialização com centróides distintos entre as execuções, ambas finalizaram com centróides idênticos, resultando nos mesmos clusters.

Execução com Seed = 10

```
Number of iterations: 16
Within cluster sum of squared errors: 158.78682155486774

Initial starting points (random):

Cluster 0: 7.7,0.49,0.26,1.9,0.062,9,31,0.9966,3.39,0.64,9.6
Cluster 1: 5.4,0.74,0,1.2,0.041,16,46,0.99258,4.01,0.59,12.5
Cluster 2: 5,1.02,0.04,1.4,0.045,41,85,0.9938,3.75,0.48,10.5
Cluster 3: 7.6,0.51,0.15,2.8,0.11,33,73,0.9955,3.17,0.63,10.2
Cluster 4: 9.8,0.34,0.39,1.4,0.066,3,7,0.9947,3.19,0.55,11.4
Cluster 5: 7,0.56,0.13,1.6,0.077,25,42,0.99629,3.34,0.59,9.2

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2          3          4          5
                          (1599.0)   (288.0)    (218.0)    (181.0)    (268.0)    (251.0)    (393.0)
=============================================================================================================
fixed acidity              8.3196     8.3733      8.422     6.3956     8.2459    11.2167     7.3097
volatile acidity           0.5278     0.5343     0.3459     0.6008     0.5264     0.4132     0.6646
citric acid                 0.271     0.2787     0.4147     0.0911     0.3084     0.5442     0.0684
residual sugar             2.5388     2.2764     2.4367     2.2602     3.2235     2.8335     2.2609
chlorides                  0.0875     0.0947     0.0737      0.068       0.09     0.1086     0.0836
free sulfur dioxide       15.8749    10.9444    13.2202    19.4309     30.319    10.8406    12.6883
total sulfur dioxide      46.4678    45.0521    30.5596    43.9503    93.8284    31.8845    34.5064
density                    0.9967     0.9972     0.9952     0.9942     0.9975     0.9988     0.9966
pH                         3.3111     3.2627     3.2829     3.4954     3.2846      3.149      3.399
sulphates                  0.6581     0.6424     0.7096     0.6341     0.6556     0.7497     0.5955
alcohol                    10.423     9.7243    11.7248    11.7955     9.8398    10.4465     9.9634

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

=== Model and evaluation on training set ===
Clustered Instances

0       288 ( 18%)
1       218 ( 14%)
2       181 ( 11%)
3       268 ( 17%)
4       251 ( 16%)
5       393 ( 25%)

Class attribute: quality
Classes to Clusters:

   0   1   2   3   4   5  <-- assigned to cluster
   0   0   0   1   2   7 | 3
  10   3   8   4   4  24 | 4
 157  16  34 179  80 215 | 5
 114 103 110  72 112 127 | 6
   6  87  25  12  49  20 | 7
   1   9   4   0   4   0 | 8

Cluster 0 <-- 4
Cluster 1 <-- 7
Cluster 2 <-- 8
Cluster 3 <-- 3
Cluster 4 <-- 6
Cluster 5 <-- 5

Incorrectly clustered instances :     1170.0   73.1707 %
```

Execução com Seed = 42

```
Number of iterations: 16
Within cluster sum of squared errors: 159.00303341277748

Initial starting points (random):

Cluster 0: 7.9,0.33,0.23,1.7,0.077,18,45,0.99625,3.29,0.65,9.3
Cluster 1: 10.2,0.645,0.36,1.8,0.053,5,14,0.9982,3.17,0.42,10
Cluster 2: 10.1,0.37,0.34,2.4,0.085,5,17,0.99683,3.17,0.65,10.6
Cluster 3: 7.8,0.52,0.25,1.9,0.081,14,38,0.9984,3.43,0.65,9
Cluster 4: 7.3,0.66,0,2,0.084,6,23,0.9983,3.61,0.96,9.9
Cluster 5: 10,0.31,0.47,2.6,0.085,14,33,0.99965,3.36,0.8,10.5

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2          3          4          5
                         (1599.0)   (284.0)    (244.0)    (219.0)    (447.0)    (185.0)    (220.0)
=============================================================================================================
fixed acidity              8.3196     8.1377     8.7484      8.416     7.3991     6.3865    11.4791
volatile acidity           0.5278     0.5292     0.5136     0.3447     0.6526     0.6025     0.4078
citric acid                 0.271     0.2942     0.3399      0.417     0.0854     0.0902     0.5484
residual sugar             2.5388     3.1486     2.3139     2.4434     2.2496     2.2514     2.9255
chlorides                  0.0875     0.0875      0.118     0.0737     0.0839     0.0678     0.0911
free sulfur dioxide       15.8749    29.6813    11.2664    13.2237    12.1085    19.3027    10.5727
total sulfur dioxide      46.4678    92.4261    43.6066    30.5479    35.2528     43.827    31.1682
density                    0.9967     0.9974     0.9975     0.9953     0.9966     0.9942      0.999
pH                         3.3111     3.2953     3.2361     3.2842     3.3807     3.4954     3.1452
sulphates                  0.6581     0.6297      0.723     0.7093     0.5908     0.6342      0.729
alcohol                    10.423     9.8027      9.784    11.7169     9.9162    11.7669    10.5439


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

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

Clustered Instances

0       284 ( 18%)
1       244 ( 15%)
2       219 ( 14%)
3       447 ( 28%)
4       185 ( 12%)
5       220 ( 14%)


Class attribute: quality
Classes to Clusters:

   0   1   2   3   4   5  <-- assigned to cluster
   1   1   0   7   0   1 | 3
   4   8   3  26   8   4 | 4
 194 125  17 247  35  63 | 5
  75  98 103 147 113 102 | 6
  10  11  87  20  25  46 | 7
   0   1   9   0   4   4 | 8

Cluster 0 <-- 3
Cluster 1 <-- 4
Cluster 2 <-- 7
Cluster 3 <-- 5
Cluster 4 <-- 6
Cluster 5 <-- 8

Incorrectly clustered instances :     1139.0   71.232  %
```

4.3. Heart Disease

Aplicação do algoritmo K-means na base Heart Disease, Cleveland, com a geração de 4 clusters.

No experimento com esta base a mudança do seed gerou a diferença mais significativa dentre os três realizados, além da execução ter sido finalizada com centróides bem diferentes, a composição dos clusters se mostrou também bem diferentes inclusive na quantidade de elementos em cada um.

Execução com Seed = 10

```
Number of iterations: 7
Within cluster sum of squared errors: 270.53338714391657

Initial starting points (random):

Cluster 0: 65,0,3,140,417,1,2,157,0,0.8,1,1,3
Cluster 1: 61,1,4,148,203,0,0,161,0,0,1,1,7
Cluster 2: 45,1,4,115,260,0,2,185,0,0,1,0,3
Cluster 3: 66,1,4,120,302,0,2,151,0,0.4,2,0,3

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3
                (303.0)     (83.0)     (67.0)     (71.0)     (82.0)
=============================================================================
age              54.4389    55.3012     54.403    50.0282    57.4146
sex               0.6799          0     0.9403          1      0.878
cp                3.1584      3.012     3.3731     2.6338     3.5854
trestbps        131.6898   130.5663    129.403    129.662   136.4512
chol            246.6931   259.3253   236.4478   235.6761   251.8171
fbs               0.1485     0.0964      0.194     0.1408     0.1707
restecg           0.9901     0.8916          0     0.9014     1.9756
thalach         149.6073   153.3855        143   163.5493   139.1098
exang             0.3267     0.1566     0.5224     0.0423     0.5854
oldpeak           1.0396     0.6265     1.2522     0.5634     1.6963
slope             1.6007     1.4699     1.7463     1.2958      1.878
ca                0.6722     0.3855     0.7862      0.357     1.1423
thal              4.7342     3.1173     6.7423     3.0423     6.1951


Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       83 ( 27%)
1       67 ( 22%)
2       71 ( 23%)
3       82 ( 27%)


Class attribute: num
Classes to Clusters:

  0  1  2  3  <-- assigned to cluster
 72 23 56 13 | 0
  7 14 12 22 | 1
  2 16  3 15 | 2
  2 12  0 21 | 3
  0  2  0 11 | 4

Cluster 0 <-- 0
Cluster 1 <-- 2
Cluster 2 <-- 1
Cluster 3 <-- 3

Incorrectly clustered instances :     182.0    60.066  %
```

Execução com Seed = 42

```
Number of iterations: 6
Within cluster sum of squared errors: 289.64838142712307

Initial starting points (random):

Cluster 0: 58,1,3,140,211,1,2,165,0,0,1,0,3
Cluster 1: 66,1,4,112,212,0,2,132,1,0.1,1,1,3
Cluster 2: 63,1,4,140,187,0,2,144,1,4,1,2,7
Cluster 3: 61,1,4,148,203,0,0,161,0,0,1,1,7

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3
               (303.0)     (71.0)     (31.0)    (100.0)    (101.0)
=============================================================================
age             54.4389    55.0986    57.0323      56.17    51.4653
sex              0.6799     0.5634     0.5806       0.87      0.604
cp               3.1584     2.7183     3.4516       3.73     2.8119
trestbps       131.6898   133.9859   130.2903     134.77   127.4554
chol           246.6931   259.3662   251.1935     248.85   234.2673
fbs              0.1485      0.169     0.0968       0.18     0.1188
restecg          0.9901     1.9859     1.3226       1.18          0
thalach        149.6073   158.2676   142.3226      137.3   157.9406
exang            0.3267          0          1       0.68          0
oldpeak          1.0396     0.6606     1.0161      1.768     0.5921
slope            1.6007     1.4789     1.7097       1.89     1.3663
ca               0.6722     0.3944     0.6774     1.1667     0.3764
thal             4.7342     3.5456          3     6.8573          4


Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       71 ( 23%)
1       31 ( 10%)
2      100 ( 33%)
3      101 ( 33%)


Class attribute: num
Classes to Clusters:

  0  1  2  3  <-- assigned to cluster
 55 16 11 82 | 0
 13  6 26 10 | 1
  1  3 25  7 | 2
  2  4 27  2 | 3
  0  2 11  0 | 4

Cluster 0 <-- 1
Cluster 1 <-- 2
Cluster 2 <-- 3
Cluster 3 <-- 0

Incorrectly clustered instances :     178.0    58.7459 %
```