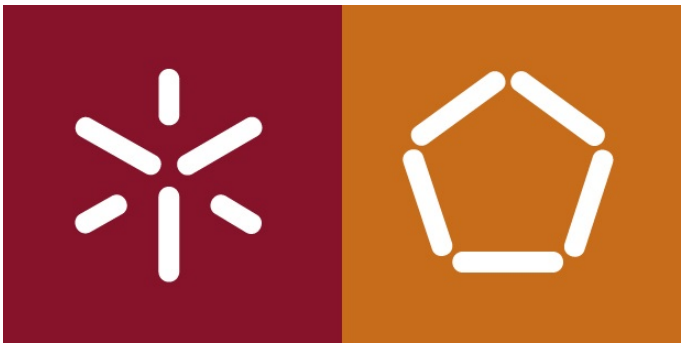

UNIVERSIDADE DO MINHO

DEPARTAMENTO DE INFORMÁTICA

DESCOBERTA DE CONHECIMENTO

Ficha 6 - Algoritmos de Classificação

5 de Abril de 2019

Gil Cunha (A77249)

Conteúdo

1 Parte 1 - Data Understanding

1.1 Exercicio 1

Atributos do *dataset heart-c.arff*:

a)

- age numeric
- sex nominal male,female
- cp nominaltyp_angina,asympt,non_anginal,atyp_angina
- trestbps numeric
- col numeric
- chol numeric
- fbs binary t,f
- restecg nominalleft_vent_hyper,normal,st_t_wave_abnormality
- thalach numeric
- exang binary no,yes
- oldpeak numeric
- slope nominal down,flat,up
- ca numeric
- thal nominal fixed_defect,normal,reversable_defect
- num binary $< 50, > 50\_1$

b)

Atributo **ca** tem 5 dados nulos (2% *missing values*).
Atributo **thal** tem 2 dados nulos (1% *missing values*).
Todos os restantes atributos não possuem *missing values*.

c)

Apenas é possível saber estas estatisticas, dos seguintes atributos numéricos:

- age - min:29 ,max:77 , mean:54.366 ,stddev:9.082 ;
- trestbps - min:94 ,max:200 , mean:131.624 ,stddev:17.538 ;
- col - min:0 ,max:1 , mean:0.275 ,stddev:0.118 ;
- chol - min:126 ,max:564 , mean:246.264 ,stddev:51.831 ;
- thalach - min:71 ,max:202 , mean:149.647 ,stddev:22.905 ;
- oldpeak - min:0 ,max:6.2 , mean:1.04 ,stddev:1.161 ;

- ca - min:0 ,max:3 , mean:0.674 ,stddev:0.938 ;

d)

- age :4 (1%) valores únicos;
- trestbps :16 (5%) valores únicos;
- col :62 (20%) valores únicos;
- chol :62 (20%) valores únicos;
- fbs :28 (9%) valores únicos;
- exang: 10 (3%) valores únicos

e)

Ao passar o cursor do rato por cima de cada coluna no histograma, apresenta-se uma mensagem a identificar o valor do atributo o qual a coluna representa e o número de registos que têm esse valor para esse atributo no *dataset*.

Ao analisar cada atributo podemos deduzir que, quanto maior for a idade e oldpeak, maior é a probabilidade de ter uma doença no coração. Também podemos deduzir que pessoas do sexo masculino têm maior probalidade de ter este tipo de doenças.Registos que contêm os valores reversable_defect no atributo thal, flat no atributo slope e asympt no atributo cp, têm maior probabilidade de ter doenlas no coração.

f) Atributos com *outliers*: trestbps, col, chol, oldpeak.

1.2 Exercicio 2

a) Os atributos mais ligados à probabilidade de ter doenças de coração é a age e thal.

b) Correlação positiva: age- trestbps.
Correlação negativa: talach-oldpeak, talach-age.

1.3 Exercicio 3

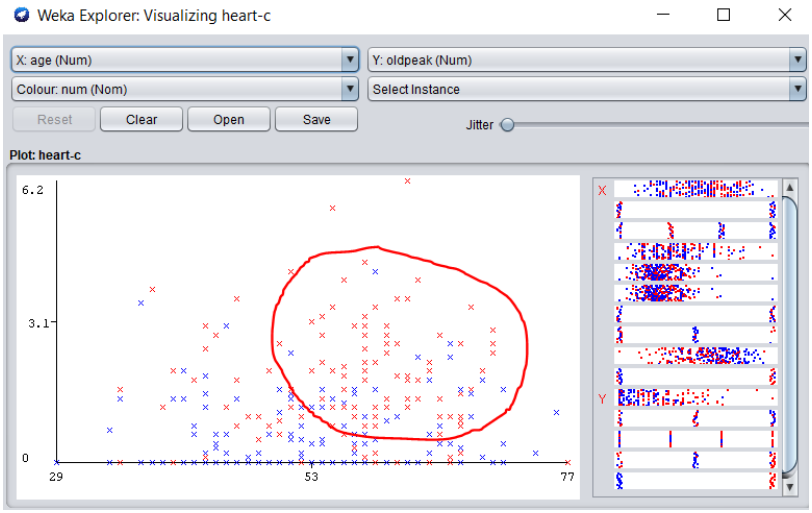

Figura 1: age-oldpeak

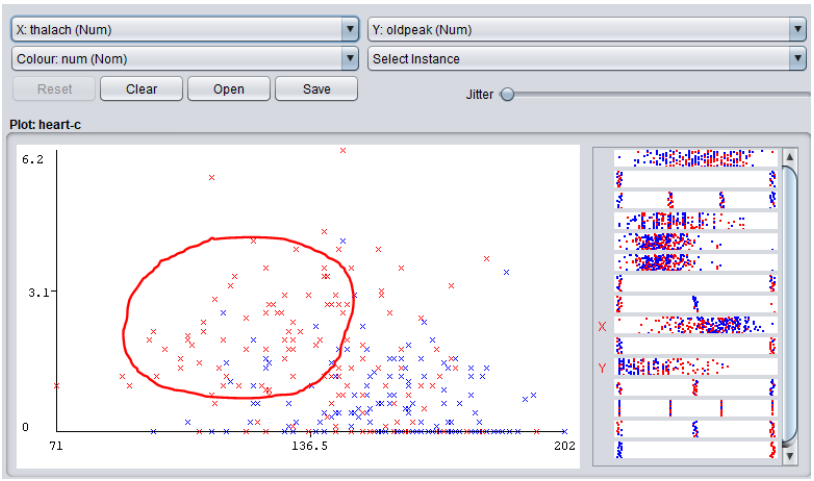

Figura 2: thalach-oldpeak

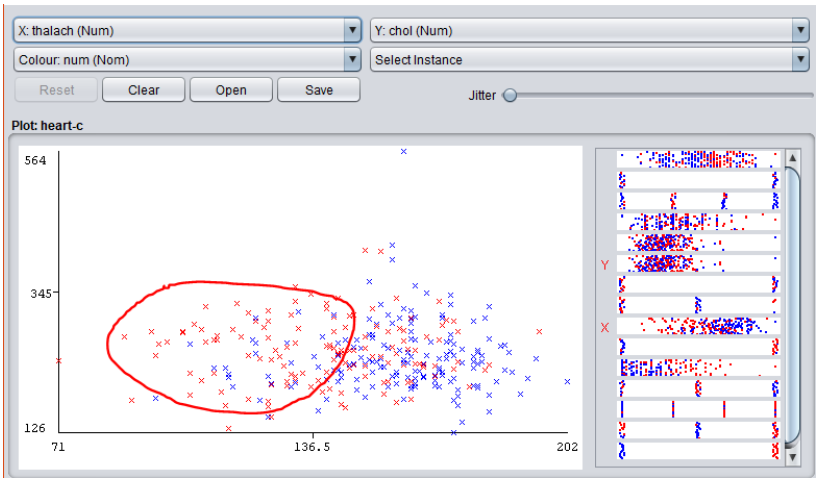

Figura 3: thalach-chol

2 Parte 2 - Data Preprocessing

2.1 Exercicio 1

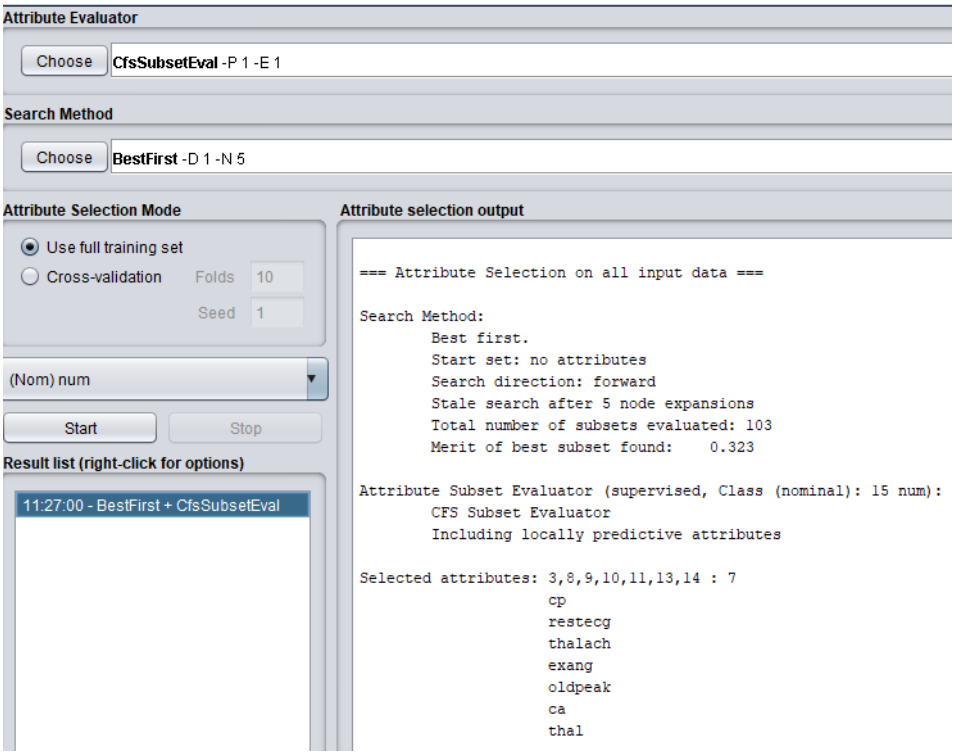

Figura 4: O filtro **AttributeSelection = BestFirst + CfsSubsetEval** selecionou os atributos *cp, restecg, thalach, exang, oldpeak, ca, thal e num*, como atributos com capacidade para desenvolver bons modelos de previsão, de entre todos os atributos utilizados no dataset.

Apesar de estes atributos terem sido selecionados, adicionei tambem o atributo ***age*** visto que, após uma análise prévia, identifiquei este atributo com uma boa relação com o atributo classe *num*.

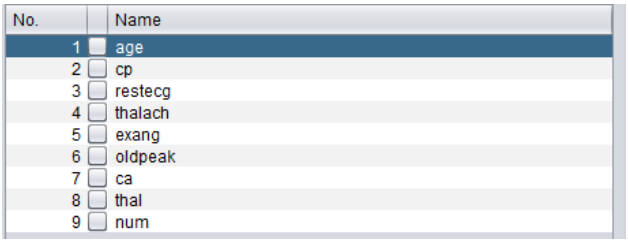

Figura 5: Seleção de atributos

2.2 Exercicio 2

Os valores em falta do atributo **ca** foram substituidos pela média: 0.7;
Os valores em falta do atributo **thal** foram substituidos pelo atributo mais frequente: normal;

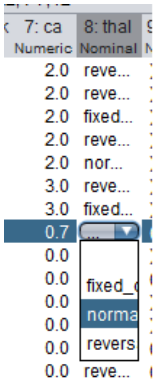

Figura 6: Tratamento de valores nulos

3 Parte 3 - Mining the Data

3.1 Exercicio 1

É possível observar que com o dataset processado obtemos melhores resultados, nomeadamente, uma percentagem de instâncias corretamente classificadas maior.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         217               71.6172 %
Incorrectly Classified Instances        86               28.3828 %
Kappa statistic                          0.4305
Mean absolute error                      0.2838
Root mean squared error                  0.5328
Relative absolute error                 57.2125 %
Root relative squared error            106.9685 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,715    0,283    0,752      0,715    0,733      0,431    0,716     0,693     <50
                 0,717    0,285    0,678      0,717    0,697      0,431    0,716     0,615     >50_1
Weighted Avg.    0,716    0,284    0,718      0,716    0,717      0,431    0,716     0,657

=== Confusion Matrix ===

   a   b   <-- classified as
 118  47 |   a = <50
  39  99 |   b = >50_1
```

Figura 7: OneR com dataset original - cross-validation 10 folds

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         219               72.2772 %
Incorrectly Classified Instances        84               27.7228 %
Kappa statistic                          0.4424
Mean absolute error                      0.2772
Root mean squared error                  0.5265
Relative absolute error                 55.882  %
Root relative squared error            105.7174 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,733    0,290    0,752      0,733    0,742      0,443    0,722     0,696     <50
                 0,710    0,267    0,690      0,710    0,700      0,443    0,722     0,622     >50_1
Weighted Avg.    0,723    0,279    0,724      0,723    0,723      0,443    0,722     0,663

=== Confusion Matrix ===

   a   b   <-- classified as
 121  44 |   a = <50
  40  98 |   b = >50_1
```

Figura 8: OneR com dataset processado - cross-validation 10 folds

Quando é utilizado o prório dataset de treino, como dataset de teste, é de esperar que a percentagem de instâncias classificadas corretamente seja ainda maior, visto que está a testar casos, com os quais aprendeu anteriormente. Isto é um caso de ***overfitting***. Se testado com casos novos (como visto nas imagens anteriores), os resultados apresentados são mais baixos.

```
=== Summary ===

Correctly Classified Instances         233               76.8977 %
Incorrectly Classified Instances        70               23.1023 %
Kappa statistic                          0.5331
Mean absolute error                      0.231
Root mean squared error                  0.4806
Relative absolute error                 46.572  %
Root relative squared error             96.5137 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,800    0,268    0,781      0,800    0,790      0,533    0,766     0,734     <50
                 0,732    0,200    0,754      0,732    0,743      0,533    0,766     0,674     >50_1
Weighted Avg.    0,769    0,237    0,769      0,769    0,769      0,533    0,766     0,706

=== Confusion Matrix ===

   a   b   <-- classified as
 132  33 |   a = <50
  37 101 |   b = >50_1
```

Figura 9: OneR com dataset processado - training set

3.2 Exercicio 2

Pelas imagens seguintes podemos verificar que, ao aplicar JRip ao dataset original, com pruning, obtemos melhores resultados (com cross-validation 10 fold), pois é mais geral: o facto de ter mais atributos e menos ramos na arvore de decisão (pruning), torna a árvore mais geral e indicada para avaliar novos casos de teste.

```
JRIP rules:
===========

(cp = asympt) and (ca >= 1) => num=>50_1 (77.0/5.0)
(thal = reversable_defect) and (thalach <= 142) => num=>50_1 (24.0/3.0)
(thal = reversable_defect) and (restecg = left_vent_hyper) => num=>50_1 (23.0/8.0)
(oldpeak >= 2.6) => num=>50_1 (9.0/3.0)
 => num=<50 (170.0/24.0)

Number of Rules : 5


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         246               81.1881 %
Incorrectly Classified Instances        57               18.8119 %
Kappa statistic                          0.6183
Mean absolute error                      0.2651
Root mean squared error                  0.3841
Relative absolute error                 53.4387 %
Root relative squared error             77.1287 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,861    0,246    0,807      0,861    0,833      0,620    0,835     0,801     <50
                 0,754    0,139    0,819      0,754    0,785      0,620    0,835     0,843     >50_1
Weighted Avg.    0,812    0,198    0,812      0,812    0,811      0,620    0,835     0,820

=== Confusion Matrix ===

   a   b   <-- classified as
 142  23 |   a = <50
  34 104 |   b = >50_1
```

Figura 10: JRip com dataset original - 10X - pruning

```
JRIP rules:
===========

(cp = asympt) and (ca >= 0.7) => num=>50_1 (78.0/5.0)
(thal = reversable_defect) and (thalach <= 141) => num=>50_1 (23.0/3.0)
(thal = reversable_defect) and (cp = asympt) => num=>50_1 (12.0/3.0)
(age >= 57) and (age <= 61) and (ca >= 1) => num=>50_1 (11.0/2.0)
 => num=<50 (179.0/27.0)

Number of Rules : 5


Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         225               74.2574 %
Incorrectly Classified Instances        78               25.7426 %
Kappa statistic                          0.4798
Mean absolute error                      0.3139
Root mean squared error                  0.4429
Relative absolute error                 63.2836 %
Root relative squared error             88.9291 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,776    0,297    0,757      0,776    0,766      0,480    0,779     0,738     <50
                 0,703    0,224    0,724      0,703    0,713      0,480    0,779     0,760     >50_1
Weighted Avg.    0,743    0,264    0,742      0,743    0,742      0,480    0,779     0,748

=== Confusion Matrix ===

   a   b   <-- classified as
 128  37 |   a = <50
  41  97 |   b = >50_1
```

Figura 11: JRip com dataset processado - 10X - pruning

```
JRIP rules:
===========

(cp = asympt) and (ca >= 1) and (oldpeak >= 0.6) and (col <= 0.394977) => num=>50_1 (50.0/0.0)
(thal = reversable_defect) and (cp = asympt) and (oldpeak >= 0.8) => num=>50_1 (23.0/0.0)
(ca >= 1) and (sex = male) and (cp = asympt) and (thalach >= 150) => num=>50_1 (11.0/0.0)
(slope = flat) and (ca >= 1) and (age <= 61) and (age >= 45) => num=>50_1 (10.0/0.0)
(thal = reversable_defect) and (col >= 0.262557) and (trestbps >= 124) and (col <= 0.369863) and (age <= 57) => num=>50_1 (6.0/0.0)
(age >= 58) and (sex = male) and (col >= 0.335616) and (age <= 65) => num=>50_1 (8.0/0.0)
(oldpeak >= 2.8) and (thalach <= 144) => num=>50_1 (3.0/0.0)
(oldpeak >= 0.8) and (slope = flat) and (thalach <= 156) and (trestbps >= 128) and (thalach >= 147) and (col <= 0.3379) => num=>50_1
(thalach <= 136) and (ca >= 3) => num=>50_1 (3.0/0.0)
(thal = reversable_defect) and (age <= 48) and (slope = flat) => num=>50_1 (2.0/0.0)
(ca >= 1) and (age <= 59) and (age >= 58) => num=>50_1 (3.0/0.0)
(cp = asympt) and (age <= 42) and (thal = reversable_defect) => num=>50_1 (2.0/0.0)
(thalach <= 146) and (exang = yes) and (oldpeak <= 0.1) and (age >= 59) => num=>50_1 (3.0/0.0)
(thalach <= 97) and (age >= 62) => num=>50_1 (2.0/0.0)
 => num=<50 (171.0/6.0)

Number of Rules : 15


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         235               77.5578 %
Incorrectly Classified Instances        68               22.4422 %
Kappa statistic                          0.5421
Mean absolute error                      0.2357
Root mean squared error                  0.4645
Relative absolute error                 47.5198 %
Root relative squared error             93.2732 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,855    0,319    0,762      0,855    0,806      0,547    0,762     0,719     <50
                 0,681    0,145    0,797      0,681    0,734      0,547    0,762     0,717     >50_1
Weighted Avg.    0,776    0,240    0,778      0,776    0,773      0,547    0,762     0,718
```

Figura 12: JRip com dataset original - 10X - no pruning

```
JRIP rules:
===========

(cp = asympt) and (ca >= 0.7) and (oldpeak >= 0.6) and (age <= 63) => num=>50_1 (47.0/0.0)
(thal = reversable_defect) and (cp = asympt) and (oldpeak >= 0.8) => num=>50_1 (22.0/0.0)
(ca >= 1) and (cp = asympt) and (restecg = left_vent_hyper) and (thalach <= 158) => num=>50_1 (11.0/0.0)
(thal = reversable_defect) and (ca >= 0.7) and (restecg = left_vent_hyper) => num=>50_1 (8.0/0.0)
(thal = reversable_defect) and (age <= 50) and (age >= 48) => num=>50_1 (4.0/0.0)
(age >= 57) and (ca >= 1) and (age <= 61) and (thalach <= 156) => num=>50_1 (4.0/0.0)
(exang = yes) and (oldpeak >= 1.6) and (restecg = normal) => num=>50_1 (5.0/0.0)
(age >= 57) and (oldpeak <= 0.3) and (age <= 61) and (thalach <= 162) and (thal = reversable_defect) => num=>50_1 (5.0/0.0)
(exang = yes) and (age >= 66) and (age <= 68) => num=>50_1 (2.0/0.0)
(cp = asympt) and (exang = yes) and (age >= 59) and (age <= 63) => num=>50_1 (3.0/0.0)
(ca >= 1) and (thalach >= 160) and (cp = asympt) => num=>50_1 (5.0/0.0)
(age >= 55) and (restecg = left_vent_hyper) and (thalach >= 164) and (ca >= 1) => num=>50_1 (3.0/0.0)
(thalach <= 158) and (oldpeak >= 3) and (cp = non_anginal) => num=>50_1 (2.0/0.0)
(thalach <= 132) and (oldpeak >= 0.8) and (oldpeak <= 1.2) and (restecg = normal) => num=>50_1 (3.0/0.0)
 => num=<50 (179.0/14.0)

Number of Rules : 15


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         232               76.5677 %
Incorrectly Classified Instances        71               23.4323 %
Kappa statistic                          0.5228
Mean absolute error                      0.2571
Root mean squared error                  0.4667
Relative absolute error                 51.8205 %
Root relative squared error             93.7055 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,836    0,319    0,758      0,836    0,795      0,526    0,750     0,704     <50
                 0,681    0,164    0,777      0,681    0,726      0,526    0,750     0,699     >50_1
Weighted Avg.    0,766    0,248    0,767      0,766    0,764      0,526    0,750     0,702
```

Figura 13: JRip com dataset processado - 10X - no pruning

3.3 Exercicio 2

Os casos em que se aplicou *pruning* apresentam melhores resultados visto que tornam a árvore mais geral, podendo o caso em que o número minimo de folha é 5 apresentar uma percentagem de instâncias corretamente classificadas ligeiramente maior, visto que com mais folhas abrage-se mais casos de teste específicos.

```
=== Summary ===

Correctly Classified Instances         234               77.2277 %
Incorrectly Classified Instances        69               22.7723 %
Kappa statistic                          0.5379
Mean absolute error                      0.277
Root mean squared error                  0.4344
Relative absolute error                 55.8332 %
Root relative squared error             87.2107 %
Total Number of Instances              303
```

Figura 14: Dataset processado - 10X - minLeaves: 2 - pruning

```
=== Summary ===

Correctly Classified Instances         235               77.5578 %
Incorrectly Classified Instances        68               22.4422 %
Kappa statistic                          0.5465
Mean absolute error                      0.244
Root mean squared error                  0.4301
Relative absolute error                 49.1936 %
Root relative squared error             86.3545 %
Total Number of Instances              303
```

Figura 15: Dataset processado - 10X - minLeaves: 5 - pruning

```
=== Summary ===

Correctly Classified Instances         231               76.2376 %
Incorrectly Classified Instances        72               23.7624 %
Kappa statistic                          0.5198
Mean absolute error                      0.2924
Root mean squared error                  0.4204
Relative absolute error                 58.9493 %
Root relative squared error             84.4177 %
Total Number of Instances              303
```

Figura 16: Dom dataset original - 10X - minLeaves: 2 - no pruning

```
=== Summary ===

Correctly Classified Instances         229               75.5776 %
Incorrectly Classified Instances        74               24.4224 %
Kappa statistic                          0.5065
Mean absolute error                      0.2818
Root mean squared error                  0.4241
Relative absolute error                 56.8127 %
Root relative squared error             85.1539 %
Total Number of Instances              303
```

Figura 17: Dataset processado - 10X - minLeaves: 5 - no pruning

4 Parte 4 - Clustering Tendency

Na seguinte imagem apresentam-se os resultados do algoritmo *SimpleKMeans*, em que o número de clusters (k) é 5. Quanto menor o *squared error* melhor.

```
Final cluster centroids:
                                        Cluster#
Attribute                Full Data              0              1              2              3              4
                           (303.0)        (50.0)         (47.0)         (68.0)         (98.0)         (40.0)
=============================================================================================================
age                       54.3663         50.78        56.9574          56.25        55.8673        48.925
cp                         asympt   non_anginal         asympt   non_anginal         asympt   atyp_angina
restecg                    normal        normal         normal   left_vent_hyper   left_vent_hyper        normal
thalach                  149.6469        157.94       147.5532         156.1176        134.9592       166.725
exang                          no            no             no             no             yes            no
oldpeak                    1.0396         0.666          1.034         0.7529         1.7449        0.2725
ca                         0.6772         0.422         0.8723         0.4118         1.0582         0.285
thal                       normal        normal         normal         normal reversable_defect        normal



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

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

Clustered Instances

0       50 ( 17%)
1       47 ( 16%)
2       68 ( 22%)
3       98 ( 32%)
4       40 ( 13%)


Class attribute: num
Classes to Clusters:

  0  1  2  3  4  <-- assigned to cluster
 42 23 52 12 36 | <50
  8 24 16 86  4 | >50_1

Cluster 0 <-- No class
Cluster 1 <-- No class
Cluster 2 <-- <50
Cluster 3 <-- >50_1
Cluster 4 <-- No class
```

Figura 18: SimpleKMean - k = numClusters = 5

Podemos aplicar o algoritmos EM, para tentar descobrir um k ótimo, já que por defeito este algoritmo tenta calcular os melhores resultados tentandod descobrir um número de clusters ótimos para tal. Quanto amis negativa for a *loglikelihood* melhor.

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

Clustered Instances

0      143 ( 47%)
1       45 ( 15%)
2       48 ( 16%)
3       67 ( 22%)


Log likelihood: -12.43524


Class attribute: num
Classes to Clusters:

   0   1   2   3  <-- assigned to cluster
 119   8  32   6 | <50
  24  37  16  61 | >50_1

Cluster 0 <-- <50
Cluster 1 <-- No class
Cluster 2 <-- No class
Cluster 3 <-- >50_1

Incorrectly clustered instances :       123.0    40.5941 %
```

Figura 19: EM - k = numClusters = 4

Porém, visto que só há dois tipos de valores no atributo classe (num), o mais lógico seria apenas haver 2 clusters. No entanto, o EM determinou 4 clusters, sendo 2 para os dois valores da classe possíveis (50 e 50_1) e outros dois para onde são atribuídos os casos menos específicos e mais difíceis de determinar (possivelmente outlier

```
                                          Cluster#
Attribute                 Full Data             0                1
                            (303.0)       (181.0)          (122.0)
=================================================================================
age                         54.3663       52.7901          56.7049
cp                           asympt    non_anginal           asympt
restecg                      normal        normal   left_vent_hyper
thalach                     149.6469      157.5691         137.8934
exang                             no            no              yes
oldpeak                      1.0396        0.5994           1.6926
ca                           0.6772        0.3674           1.1369
thal                         normal        normal reversable_defect



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

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

Clustered Instances

0      181 ( 60%)
1      122 ( 40%)


Class attribute: num
Classes to Clusters:

   0   1  <-- assigned to cluster
 145  20 | <50
  36 102 | >50_1

Cluster 0 <-- <50
Cluster 1 <-- >50_1

Incorrectly clustered instances :       56.0     18.4818 %
```

Figura 20: SimpleKMean - k = numClusters = 2

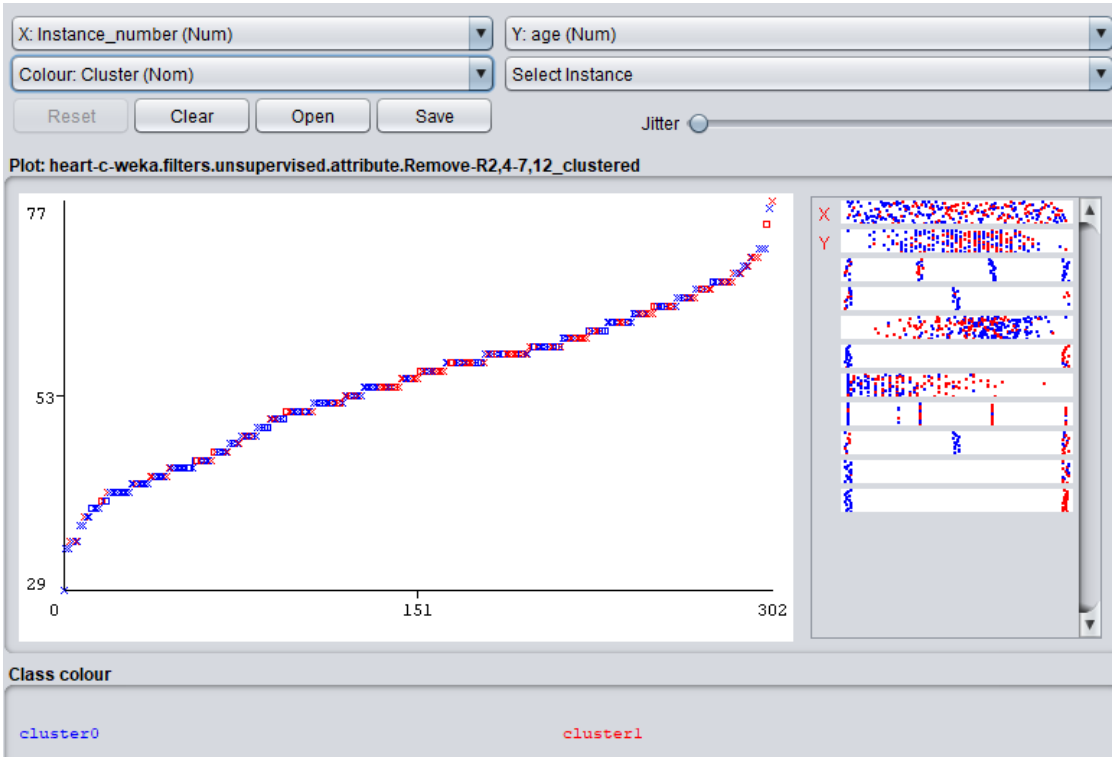

Figura 21: Visualize clusters: Cluster 0 – 50; Cluster 1 – 50_1